This video for focuses on some basic statistics facts that we will need for this class. Before we begin this video, I want to have a little bit more of a comment on the use of the right number of digits to present. For this course, I am assuming that you’ve already read the document “Some Important Points About Significant Figures” on Perusall. For the problems in class and on exams, the three-digit rule discussed in that document is what we’ll go with. For homework on mastering physics, however, answers will be judged as correct if they are within two percent of the correct value, so my suggestion is to just put in plenty of digits. You may get a message that looks something like this. If you do, don’t worry about it, the problem has been graded correctly. Mastering physics is just telling you that you did your significant figures wrong. However, since I don’t care about significant figures, it’s not really a big deal; you will get full credit either way.

So, what are the objectives for this video? We have discussed how many digits to present in class on exams and in homework, but what will we do for the laboratories that we will do in class? For labs, we will use what people in research actually use: the ideas of mean and standard deviation. Before we talk about mean and standard deviation, we have to discuss a little bit about measurement. Most objects have some variation. People come in a variety of heights, for example, and even manufactured objects like, say pencils, will have a variety of lengths. If you can measure them precisely enough, this variation may be very small, if for objects made by machines, but it will still be there. Another example that’s not just lengths or heights is the number of blood cells passing through a capillary per second. This quantity will, of course, vary from second to second. Another example might be that if you have a spring based launcher of a ball, the ball will travel slightly different distances each time, if for no other reason than the spring coils slightly different ways on the molecular level with each launch. These types of variation are intrinsic, and result in variation in your measurement. However, sometimes measuring something directly is tough, and you need to use indirect methods, like we do for our library lab. One way to get a feel for the precision of your method is to make the measurement with a few different methods, with what we expect to have similar levels of precision. The variation in the results of the different methods can give a sense of the precision of your measurements; this is how we will evaluate our methods for the library lab.

To talk about the ideas of mean and standard deviation, it’s helpful to have an example. Say we took the height of 20 men from the United States and presented the data in the table below.

|  |  |
| --- | --- |
| **Person** | **Height [ cm]** |
| 1 | 177.7 |
| 2 | 181.4 |
| 3 | 179.4 |
| 4 | 164.9 |
| 5 | 180.4 |
| 6 | 174.0 |
| 7 | 178.6 |
| 8 | 176.1 |
| 9 | 181.9 |
| 10 | 179.7 |
| 11 | 175.8 |
| 12 | 175.0 |
| 13 | 180.9 |
| 14 | 181.9 |
| 15 | 181.0 |
| 16 | 180.5 |
| 17 | 169.2 |
| 18 | 173.3 |
| 19 | 171.8 |
| 20 | 176.5 |

Note that there are no uncertainties listed in this table. Yes, the ruler that we’re using has some limit of precision, which is apparently .1 cm according to this table, but the variation between measurements is much larger than this, so the precision of the ruler won’t be too important. In a well-designed experiment the precision of your instruments should be much less than the intrinsic variation that you are trying to measure. The most complete way to report this data would be to report the entire table as we’ve done here. However, this becomes impractical as more and more data are collected. Moreover, it becomes very difficult to see trends when you just got big lists of numbers. Therefore, we need ways to characterize our data. If you’re looking for a way to characterize the data, the first thing that you might think of to do would be to take the average. Now in mathematics, the word average is replaced with the word mean; they’re synonyms. There are many different symbols for mean and each discipline seems to have their own one, so I’m going to present you with all of them. I wish we could agree on which symbol to use, but we can’t, so I’m just going to show you all of what’s out there. The Greek letter here is a very general symbol for mean. A general tip I would have is that you learn your Greek alphabet. Another way to represent mean is, let’s say we’re using the variable to represent the height of a man, then you might see or to represent the mean. The formula for the mean is given by

Many of the equations that you might see in this video can get pretty ugly looking, but they are manageable if you stop and parse them down and read what the equation is trying to tell you. A good tip for questions is to read actually right to left. So, let’s give that a shot with this equation. The letter represents an index over the measurements. Here we have 20 measurements, so is an integer that runs from 1 to 20. is one specific measurement, so is the height of the fifth person which, according to our table, is 180.4 cm. To calculate the mean, we add up all the measurements, and then divide by the number of measurements. Let’s calculate the mean. For these data, when we add up all of the measurements, we get a sum of 3540 cm. We divide by the number of measurements; take , which gives us a result of 177 cm. This is our average. For our mean: the mean provides a great starting point for characterizing data, but it’s insufficient because it’s missing a key feature. Just representing the mean gives us no clue on how spread out these data are. Phrased differently, we don’t have any information on what is the average distance for a random data point to the mean. So if we’re looking at this question, let’s try to translate this question into mathematics. Well, the distance from a given data point to the mean would be , and the average distance would be, well, take all of these different distances, , add them all up and divide by the number of measurements.

However, this idea has a problem. Some distances are lower than the mean. For example, person 6 is slightly shorter than our average. So, his distance to the mean will be positive, while some people are taller than the average, for example person 2, so their distance to the mean will be negative. If I add up positive numbers and negative numbers, I’ll probably get a result that’s very close to zero due to the cancellation. So how can we get around this problem? Well, you might think absolute values, but for calculus reasons, absolute values have some problems, so a better way to get around having negative numbers is to look at squaring them, because no matter what, when I take a number and square it, the result is positive. So, let’s look at the average squared distance from the mean. Mathematically, the average squared distance from the mean would be, take the distance from the data point to the mean, just mean minus data point, square it, add them all up and divide by the number of measurements. Now all the numbers being added are positive, so there’s no cancellation. This quantity is called the variance, and we will label it with the variable , for reasons that will become apparent in a moment. Let’s calculate the variance for this data. Again, variance is kind of an ugly formula, so you really got to slow down and take it one piece at a time. So let’s take an entry I equals 1 and what do we do, take the entry and subtract it from the mean. This for I equals 1 is negative .7 cm. We repeat this for all of our data. We get these results. Next in variance, we see we should square so for I equals 1 the result is .49 cm2. I want to point out that we’ve now moved from cm to cm2, because when you square a number with units, you got to square the units too, and when we repeat it for all of our data and get these results to calculate variance, we take all of these numbers and add them up, which gives us 403.5 cm2. To get the variance divided by the number of measurements, which in this case is 20 giving us a variance of 20.2 cm2.

Now variance has different units than mean, as we’ve already seen. The mean for this data set is 177.0 cm while the variance is 20.2 cm2. It’s very difficult to compare numbers with different units, so to deal with this, we, instead of looking at the variance, look at the standard deviation, which we represent by. So, the standard deviation is the square root of the variance. This is why we represent variance with a . In this example to get the standard deviation, we take square root of the variance so the square root of 20.2 cm2, to give us 4.49 cm. Now we have two quantities that are both in cm, and allows us to compare them. So how do we report numbers in the laboratory exercises? In this class, most of the labs in this course will have multiple measurements. We can use these different trials to calculate a mean and a standard deviation, and we can use this standard deviation as an uncertainty and use it to tell us how many decimals we should record. In our height example, we had a mean of 177 cm and a standard deviation a 4.49 cm. An appropriate way to represent this result would be 177 plus or minus 4.5 cm. This representation has a lot of advantages; it represents the average 177, and we have the standard deviation, which gives the person reading the number a sense of the spread of the data, and we have a reasonable number of digits. I’ve gone with one digit past the decimal point, and I did this based upon the standard deviation. While our standard deviation is officially 4.49, I rounded it to 4.5, because .01 is very small relative to our standard deviation, so I can’t really trust that .01. So while I removed some certainty of nice hard sig-fig rules, this is how numbers are actually reported in research, and this is how we’ll do it in class. Part of the point of the laboratory exercises to get you some experience with this sort of usage.