Quantitative Methods in Chemistry

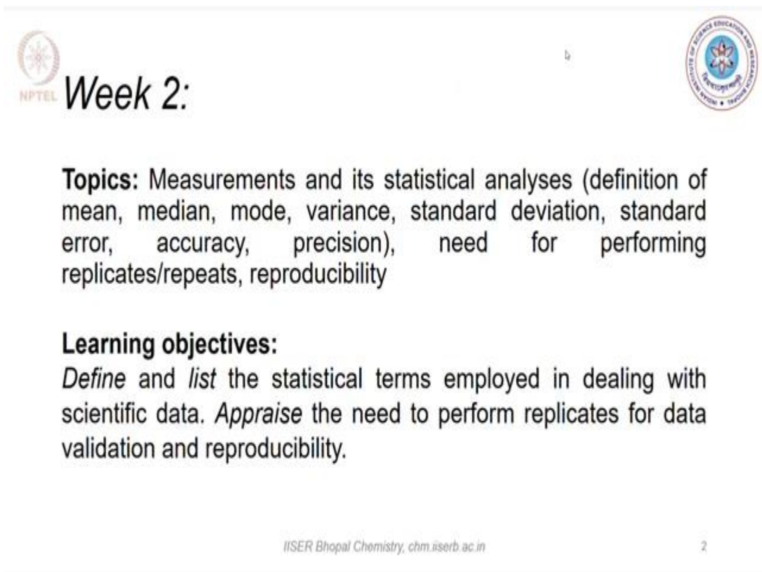
Dr. Aasheesh Srivastava
Dr. Bharathwaj Sathyamoorthy
Department of Chemistry
Indian Institute of Science Education and Research – Bhopal

Lecture – 06

Brief Introduction to Normal Distribution and Statistical Analysis

Welcome back to quantitative methods in chemistry. We are entering into the second week to start with let us take a look at what are all the learning objectives for this week.

(Refer Slide Time: 00:36)



In this week you will try to understand what is a measurement and what all statistical analysis could be performed with the data or the measurements that are being obtained. We would start by defining what is mean? What is median? What is mode? Variance, standard deviation, standard error, accuracy and precision. These are all the different statistical tools that we will end up using to have an assessment of what kind of data or measurement that we have made.

And also we will try to understand what is the need to perform replicates, repeats and what does the word reproducibility mean in the field of data sciences. So moving ahead let us clearly define what is that we are going to do like the way we did last week? Here will be defining and listing these statistical terms. We will be employing them in dealing with scientific data to start with we will be looking at a simulated data.

And as we go forward, we will be taking examples from analytical chemistry to understand how are these terms would help us understand? And most importantly apprise the need of performing duplicates. So this would help you understand how to deal with scientific data.

(Refer Slide Time: 01:47)

So to start with let us start maybe with a little bit of philosophy. Science is basically pursued as a systematic evaluation of observations acquired from experiments that help you evaluate and at times develop and underlying hypothesis. A series of coherent experiments are performed to uncover the truth. So basically the truth is what we are trying to pursue here and this truth is pursued by performing experiments.

So let me repeat a series of coherent experiments are performed to uncover the truth that help and validating, supplementing or rebutting and existing theory. So what do we mean by this? Maybe a theory already exists for instance Newtons laws of motion you could get some data to validate it meaning that you perform experiments that tend to show all the data that is acquired the new experiments agree with the existing theory or it could be used towards supplementing an existing theory which means that there are some missing gaps in a given theory and your data will help supplement and improvise the hypothesis towards making it even more fulfilling theory.

On the other hand rebut an existing theory meaning that you are able to perform experiments that go against the underlying hypothesis of an already existing theory and therefore the data that we tend to seek truth with might help us uncover this. Generally the truth is assumed to be obtained when similar results are obtained across many different people, many different labs and when they are repeatedly probing the same entity with complementary experiments.

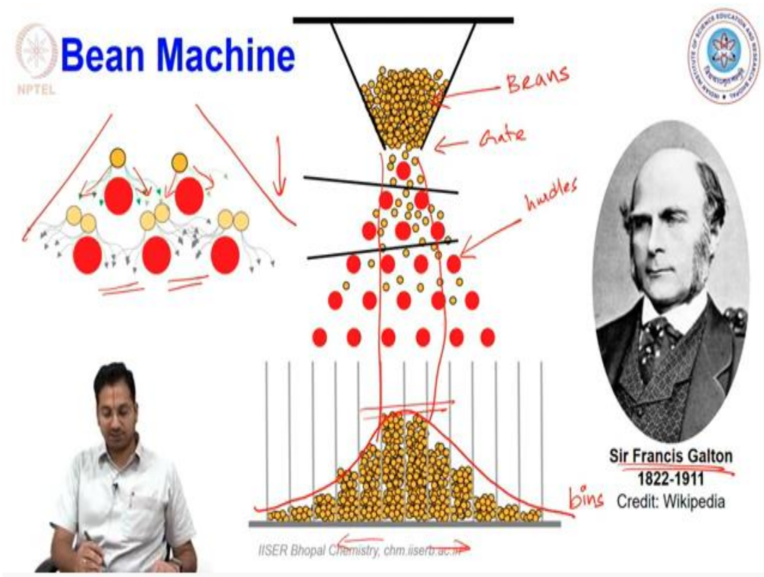
So basically the reproducibility meaning that if a given entity is measured by a multiple people by different labs at different points of time using different types of experiments yielding the same value helps you adjust the fact that you have indeed seek the truth and I have actually found the right value. For instance the mass of an electron has been measured by various different theories and its mass has been determined to be 9.1093837015 times 10 power –31 kilogram.

So you are able to realize most of the calculation we end up using 9.1 10 power -3 kilogram as the mass of an electron. But here you are able to realize to the level of precision at which the entire measurement has been made over the years and what do the values in the brackets mean they kind of mean the uncertainty associated the last 2 digits that have been shown. So this week of lectures will help you understand how to deal with numbers. And what do these numbers mean?

Remember that you entirely spent a week of understanding how to determine concentration of different chemicals in various possible units? Let us say that I am measuring concentration of sodium chloride in sea water and you are also doing the same experiment in a slightly different using a slightly different approach and then we would like to compare our results to see what is that we have gotten?

So to compare such numbers and also to understand okay to what level of precisions can we give numbers the statistical tools that we will be introducing in todays class would help you towards the same.

(Refer Slide Time: 05:41)



Before going ahead let us try to do an experiment or at least think about an experiment that was proposed by Sir Francis Galton. This is important to understand because most of the distributions that come up in science follow something called a Gaussian distribution Gaussian function and to understand and illustrate the same Sir Francis Galton came up with a very simple experiment. Let us assume that you have a bunch of beans that have been saved in a reservoir and this is the gate that might unleash them onto the bins that are present here.

So let us say that you open the gate and the beans start to fall of course due to gravity. What kind of bins will actually get these beans at the end of the experiment? So basically you one might always argue saying that okay its right at the center we are only going to have these bins occupied or the other person who might like an uniform society but guess all the bins here would be uniformly occupied.

Let us try to look at what when such an experiment is performed? What kind of results come up? So the moment the gate is open you are going to have beams that start to fall down due to gravity and as they fall down they are going to interact with the hurdles that are being kept up here and while they are interacting what is going to end up happening? Is that the beans could go on either side after their interaction with these hurdles?

Remember the probability of the beans going on either side is equally likely and if you have many such beans falling apart or falling from the top you are going to have good amount of distribution that goes both ways. But remember as you start going to additional levels at the bottom you are also going to have the beans that start to interact and that is going to push things one way or the other right.

So what is this going to end up happening? Is to make the beans as they start to fall apart start disperse them across from the initial point that they started with. So basically trying to give an idea instead of all the beans coming right at the center of where it was released. This is not going to end up happening. We are going to see what is going to happen when such an experiment is performed.

When this was done one is able to see the distribution that beautifully comes out looks like this where significant portion falls from where were the beans were released. But there are also bins that go farther away from where they were released also populated but the population keeps decreasing as we go farther away from the point of release. So this distribution you are going to quickly see is like the Gaussian function.

So we are going to call it a normal distribution or a Gaussian distribution okay. All this is in terms of a pictorial representation why do not we take an example of the real Galton board.

(Refer Slide Time: 09:11)

So the Galton board is available for sale at GaltonBoard.com and you are able to see the simulation that goes here which will be shown to you right now.

(Refer Slide Time: 09:22)

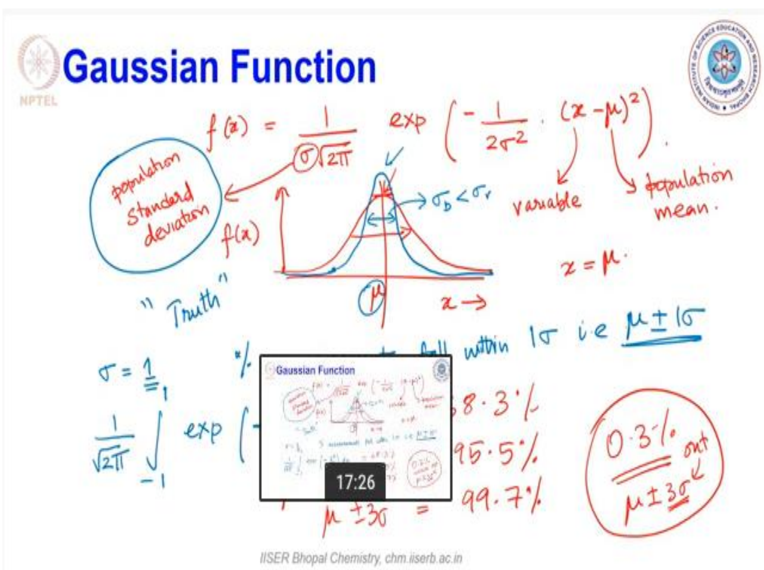
What you are able to see here is that this is a board that is made to do the simulation as we saw in the slide a moment earlier. So first step would involve getting all the beans into the reservoir as we see here. Once that is done the immediate thing is to open the gate or in this case is just to flip the board that will result in these given hurdles interacting with the beans that come about and the bins that are given here would be populated as I had shown in the slides.

So let us take a look at such a simulation. So now when you flip this you are able to see the beans start to come off from the reservoir to these different bins and you realize they get populated in different places. One thing that is important for us to note here is that from the point of release most of the bins do end up occupying that spot. But you do see that there are bins that are much farther away which ends up getting some of the beans as well.

So what happens if we repeat this experiment if you are able to see there are certain kind of distribution there. Let us repeat it again to just whether we get the same set of distribution and you realize there are subtle changes that ends up happening. So basically this is what I started the whole discussion with today. These are the kind of measurements we would end up doing in science where you keep repeating a given experiment and try to understand in this case where did the beads actually fall off from.

And you would be able to say if you take the average of all these measurements will come at the peak of this Gaussian distribution and you would say this is where the beads were released from in this Galton board.

(Refer Slide Time: 10:58)



Let us define the Gaussian function. So there are various parameters that I have been written here. Here the sigma is called as the standard deviation more precisely the population standard deviation and x is the variable and mu is the population mean or average. So now this function basically when plotted results in something that looks like this. So what one is able to observe that when x takes the value of mu you are going to have the maximum that comes up for this function.

But as x starts to go away from you since you have an e power - function that is going to fall off on either side. So this is a typical Gaussian function and the average is defined as the peak that you are at the peak or the maxima that you find for this function okay. Now what does this variable standard deviation mean? The standard deviation means how big the Gaussian distribution is.

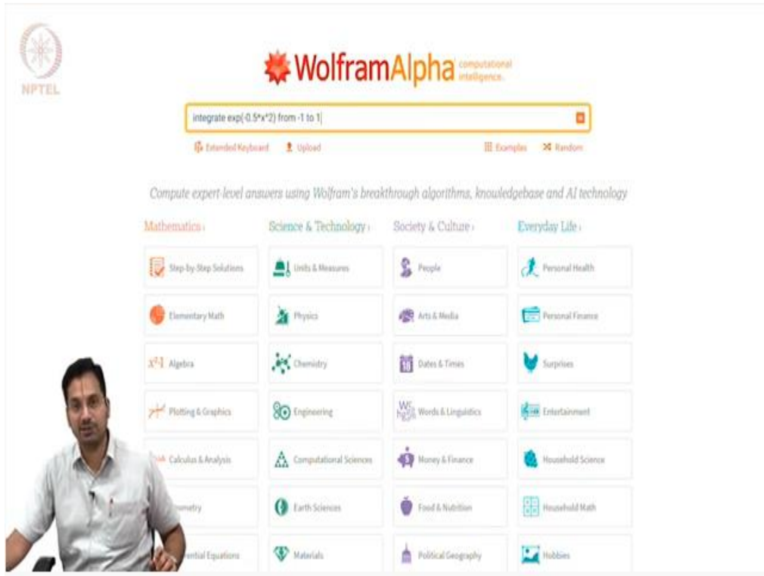
Basically you could have a similar situation where the average is the same but the width is smaller. Of course I am drawing the arrow at different places there should be somewhere here the width is smaller. So in this case the sigma of blue is less than sigma of red. This indicates the fact that the uncertainty associated with the measurement of x in terms of the blue is less than that of the red has seen from this Gaussian distribution okay.

So now we are able to realize the fact that the level of uncertainty that comes is associated with this parameter standard deviation and the truth that we are trying to seek is obtained from the average parameter of mu okay good. So now that we have seen these let us try to understand how much of this average and standard deviation go with one another. So one way of looking at it is to integrate this function.

So let us assume a sigma that is associated with the given measurement of 1 then let us ask ourselves the question how much % of measurements fall within1 standard deviation that is mu +- 1 sigma. So this can be easily done by integrating between the limits -1 to +1. For the sake of simplicity let us take the average value of truth that we are trying to seek out this experiment is 0. So that makes our math a little easy to see.

So you are going to have 1/square root of 2 pi times e power -x square/2 dx and its integrated between -1 to +1 because you are trying to find what is the value of 1 standard deviation. Let us quickly use another mathematical tool to understand the same.

(Refer Slide Time: 14:55)



So in for this example I am going to be making use of a software called WolframAlpha that is powered by Mathematica to integrate to make our life a little easy. So integrate the function e power -0.5 which is 1/2 times x power 2 from -1 to 1. So this is nothing but the Gaussian function that we have defined and of course we here we are started by saying that the uncertainty or the standard deviation associated is 1 unit which is what reduces to this.

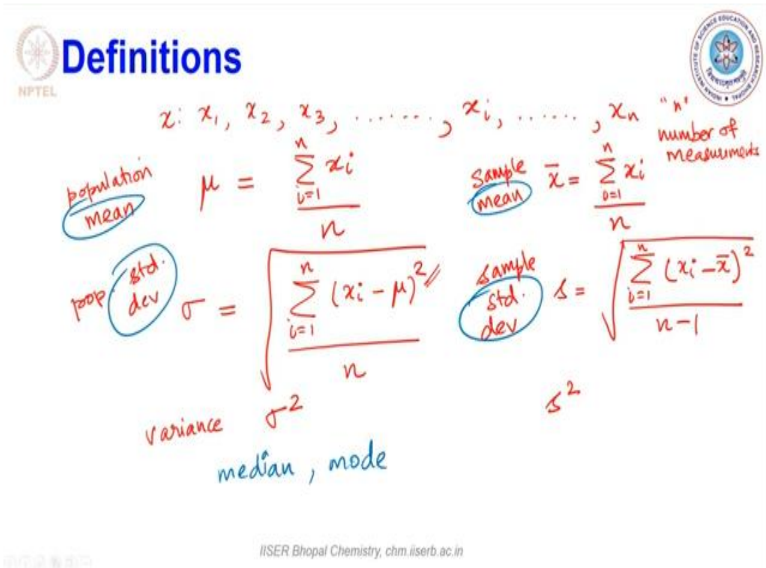
And of course there is also pre-factor that goes as 0 2 star pi to the power -0.5 because this is nothing but 1 over of that. So let us integrate it between the limits -1 to +1 to see what you get. So this is the integration you wanted to go from -1 to +1 e power -x square/2/square root of 2 pi and what you are able to see is that you get a value of 0.683. So let us do the same exercise and ask ourselves how much will it be between -2 to +2 and what do we end up getting here.

We get about 95.5% and let us go between -3 to +3 that ends up to be 99.7%. So going back to our presentation this we found to be somewhere equal to 68.3 % of course it is a 0.683 I am converting it to % between the limits mu + - 2 sigma comes up to 95.5%.

We can call it 96% and mu + - 3 sigma comes up to 99.7%. So what one is able to observe here is that there are still 0.3% of your measurements that fall outside the 6 sigma that is mu +-3 sigma that is 0.3% that falls out of this value distribution and this is a very important point that one has to understand that although your average gives you a good understanding.

Your standard deviation makes you even understand better if there is an outlier that comes into play. Almost every data point that we end up measuring in an very well setup experiment would make you understand what is the uncertainty that is associated with such a measurement.

(Refer Slide Time: 17:38)



So let us start with a few more definitions we just define what is a Gaussian function and how does the standard deviation affect the way the function looks? Basically if you have a large standard deviation the function is quite broad. On the other hand if there is a small standard deviation the function or the distribution is rather narrow and this helps you understand what is uncertainty associated with such a measurement.

So let us say you have a given number of data sets. Let us say x is the variable that you are measuring and you measured it let us say n number of times xi. This is the number of measurements that you make. Out of these number of measurements let us say you have infinitely large number of measurements then the population mean will be defined basically as the arithmetic mean where n is the total number of measurements.

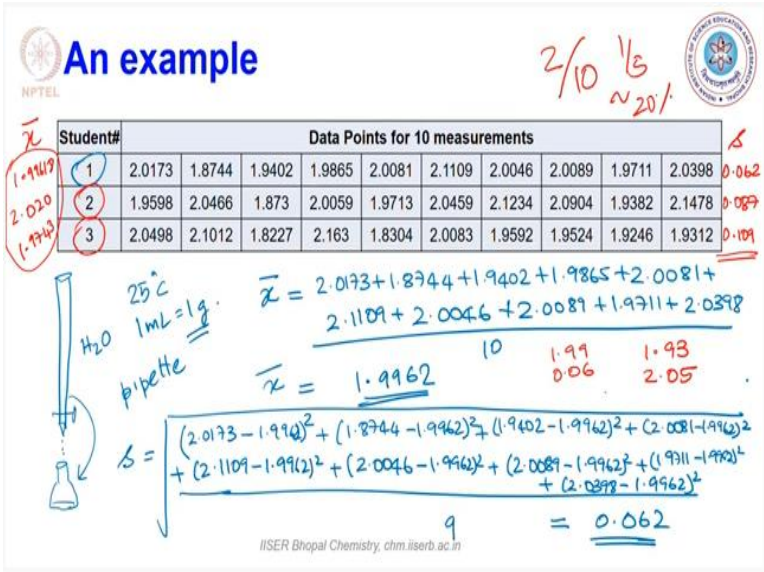
When you do not have enough number of data points we end up calling this as the sample mean which has the same formula but a slightly different symbol instead of using mu will be using x bar and the standard deviation would be the amount at which it deviates from the mean basically how far does each of the value that you are measured for this different variable x falls away from the average.

So that is going to be given by sum of from 1 to n xi - mu the whole square/n with the square root and the reason why you have a square dependence here is because you could have values of xi which is greater or smaller than the mean. It does not matter which way they are far away from you would like to quantify how far is each of the measurement xi is away from the mean that is mu.

So now when you once again do not have enough data sets this becomes called sample standard deviation. This is population standard deviation that gets a symbol s that is going to be given by sum of i = 1 to n like the last time xi of course –x bar because you change mu to x with the square the whole divided by n-1. The fact that n-1 comes in these cases is to remove the bias that exists with a small set of data set vary if we have n measurements you should be able to get the last value from the n-1 measurements that you are made.

So these are the definitions that go of course there is another definition called variance is sigma square or s square. So the value variance also helps you understand how broad is the distribution for the given set of values that you are measured for that given experiment okay. So now that we have defined all these functions there are few more that we would like to define. Of course I would like to make you guys see that we are defined mean standard deviation and median, mode and the others will be defined as we see more examples.

(Refer Slide Time: 21:29)



So let us take an example to understand how does the mean and the standard deviation go. So this is an experiment let us say a student is trying to use a burette and is trying to calibrate the bureau. How is this burette calibration going? The burette calibration is done such that where the student ends up a liquating 2 ml of solution let us say water to keep things easy the 2 ml of water is liquated into the conical flask and ends up taking this flask to weigh how much does this 2 ml wave.

So now when you have such a measurement made of course at 25 degree celsius you can assume 1 ml of water corresponds to 1 gram and if you are able to make measurements of 2 ml every time you are going to get something close to 2 grams. So this is a set of data that has been obtained by students using let us say more precise instrument to get the values and what you are able to see is that the average values are very close to 2.

You are able to realize that the student has a aliquoted quiet carefully and let us say that the student instead of using a burette use something like a micropipette. So that the values are this precise with a burette this may not be that simple or straightforward to obtain. So when such a thing is done let us determine what is the average? The average in this case is going to be for the student 1.

It is going to be the sum of the first row of elements that we have here okay. So that is going to be 2.0137+ / the total number of measurements. Since this involves 10 measurements n = 10. This is the value that you are trying to look for. Let us quickly determine what that is. So that comes up to1.99618 of course since we are giving only 4 digits this has to be 1.9962. We will be looking at significant figures as we go forward and then you have to calculate.

So this is x bar the next thing that we have to end up calculating is s which is nothing but square root of since you have only 10 measurements the denominator is going to be 9 which is n-1. This is going to be 2.0173-1.9962 the whole square. So this turns out to be 0.062 so what you are able to realize here is that the average mean value the sample mean comes up to be 1.9962 + - 0.062 and this just means that if you are having values that are within let us say 3 standard deviation in this case it is going to be 0.18 almost all the values.

Basically we are talking about 99.7% of the values from these 10 values will fall within that range why do not we take a quick look whether that is indeed the case. So what do we mean by that? So 0.18 so 1.99 + 0.18 is going to be 2.17 is the highest possible value that one could get and -1.81. So let us see whether all values here fall. Yes this falls <2 point the range >greater than 0.81 within the range, within the range, within the range, within the range, within the range.

So you are able to realize all the values fall within the 3 times standard deviation. Let us ask a question how many of these values now fall within 2 times standard deviation. So what is going to be 2 times standard deviation 2 times standard deviation for this example is going to be + - 0.12. So that is going to be 1.8722.11 and here since we are doing 2 standard deviations it is going to be 95.5% of the values.

So let us see whether all the values fall within that range yes. Yes barely makes it yes within the value, within the value, within the value barely makes it. But it is definitely within the value all these are within the value. So what you are able to realize the standard deviation helps you get an idea of whether the data points that you got gives you a very reasonable spread. Of course now let us do the last step here.

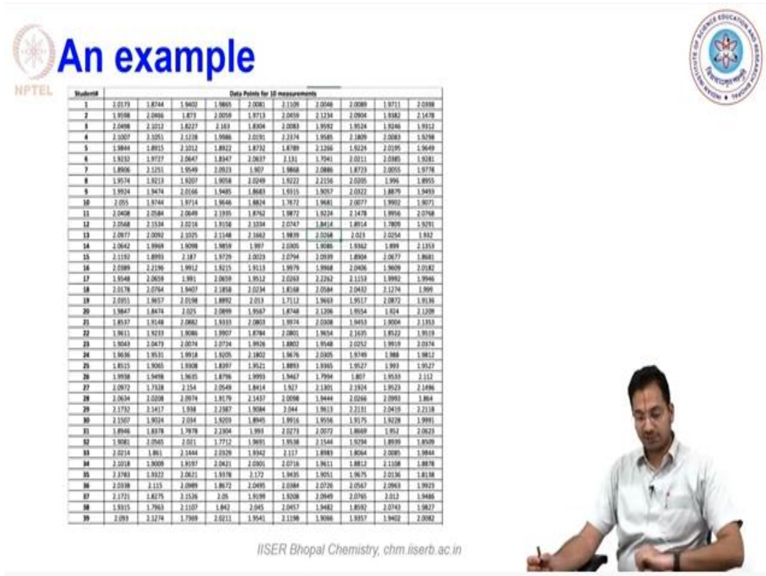
Let us start to ask how many of these values fall within 1 standard deviation. The 1 standard deviation is going to be 1.99 + - 0.06. So that is going to be 1.93 to 2.05. So right away we are able to realize whichever just scraped through last time is out of this range. This makes it, this makes it, this makes it, the next value makes it, this value does not make it, this makes it, makes it, makes it.

So what you are able to realize 2 out of 10 values which is 1 5th which amounts to over 20%. So although we said only 69% of the values meaning that about 30% could stay outside you realize that within this data said 20% stays outside and this comes up because of the fact that you do not have enough sampling to see the 30% falling out. If we had taken a little more set of data points you are going to see about definitely 30% of these values that stay out.

And now I would leave it to you guys to repeat the exercise for student number 2 and 3 and let me give you the average and standard deviation for such a case. So the answer for the average and standard deviation for these examples are let me write the average on this side and the standard deviation on the other side. The average works out to be 2.020, 1.9743 and the standard deviation as we just saw in this case is 0.062.

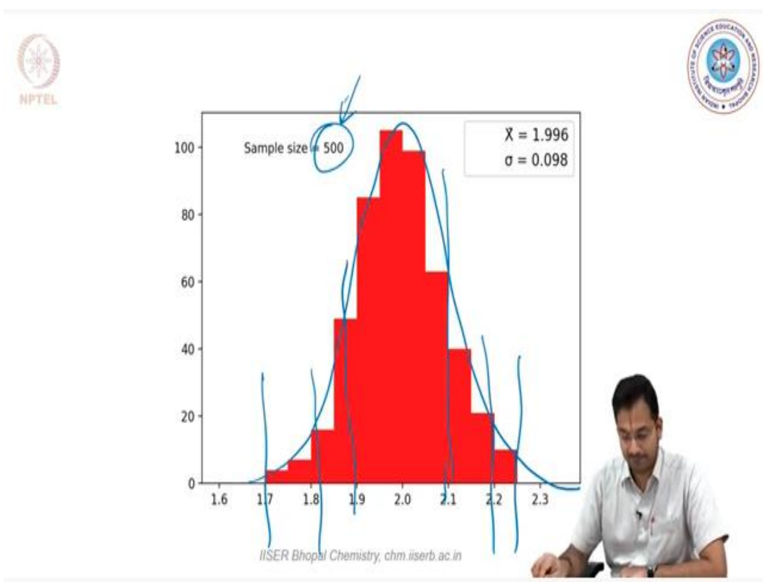
The next case is 0.087 and in the last case is 0.109. So what you are able to see when you keep on doing the measurements over and over you tend to see how far the average could change. But since the average is changing within the standard deviation of each of this measurement one could always say yeah you are actually getting comparable values across the different measurements that have been made.

(Refer Slide Time: 28:33)



Of course this does not finish the story if you want to honestly seek the true value you want to make infinitely as many measurements as possible. In the assignment you are going to get an example where we are going to give data sets across 50 students and we will try to understand what is the average and the standard deviation of course a sample mean and the sample standard deviation that has been obtained for these measurements.

(Refer Slide Time: 29:01)



And in order to facilitate for us to understand how things go better I have already started to plot it to understand how the histogram goes. This is the case where you have 10 data points and you get an average of 1.996 and a standard deviation of 0.06 and what you are able to realize here? The average stands here lot of the values are close to the average. But you are not able to see any entities that come within the 3 standard deviations meaning that which going to be 0.18 to be the triple standard deviation.

You still expect some points to exist here. But what you are able to see is that, that does not happen and most importantly you actually do not see any data sets in this given bin. So the bin that we are talking about is the width that one chooses and this for this numerical simulation we chose a bin width of 0.05 units and these are nothing with the same bins that you ended up seeing in the gallop board depending upon how big the gallop board bin is going to be?

It is going to taking more or less number of beans okay. Let us start to increase it instead of having 10 data points let us say we have 20 what ends up happening is that you starting to slowly see the other populations come in slowly and then as we keep increasing it to 30 you see that the distribution keeps changing here and there but the average has not shifted much. There has been some change in the standard deviation but within the error of the measurement this could end up being nothing significant.

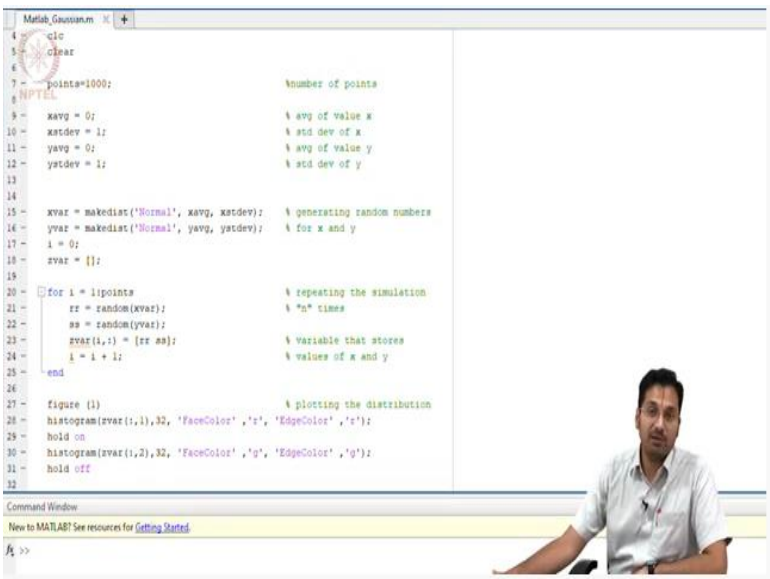
So let us keep increasing it for 40 the average still stands at 2 but you are once again seeing a weirdity that comes where in the middle of the bin things do not fall. This happens a lot of times in statistics which is where people suggest you to get as many data points as possible. So let us keep increasing the number of values you immediately saw a value that comes outside the standard deviation rather right at the edge of the standard deviation since the value changed here the standard is about 0.1.

So you are going to talk about 2+- 0.3 as the 99.7% of the values and you do see a value that pops up when you have a large number of data sets. So these things tend to happen when you are setting up an experiment. Let us keep going you slowly starting to realize that the distribution here starts to come up it was initially underrepresented. So you go forward you are able to see that it has come up properly and slowly you are getting towards a Gaussian distribution.

But as you increase the number of values you are going to realize that standard deviation also changes much lesser than before. Previously change from 0.06 to something like 0.08, 0.09 and higher. But now you are able to realize that it has saturated about at 0.09 as you increase the number of values of measurements. Let us keep doing this, let us go faster and when you have about 500 measurements.

So we had about 50 students who made 10 such measurements you are able to nicely see a distribution that comes up and you are able to realize what is the these where the standard deviations for 1 sigma and you see about 60% of the values. This will be 2 sigma where 95% of the values go. This is one where in this case you are able to see all the 100% values. But if you increase this number to a much higher number something like 5000 you will be able to see higher. This is what we are going to see in the example that follows right now.

(Refer Slide Time: 32:26)



So let us do a numerical simulation to finish off this example. In order to exemplify the point that we just learnt in terms of distribution and the measurement that goes with each other. I am going to perform a simple numerical simulation using the software MATLAB. And I am only going to briefly explain the code where all that we are trying to do is for 2 variables x and y. I will tell you why we are using 2.

So that we can compare them before and after. We are going to be generating 2 variables x and y with a given average and standard deviation which we have the power to change. And we will be seeing how the distributions change for such a simulation and we are going to be relying on this mathematical software MATLAB to generate random numbers within the average and standard deviation assuming a Gaussian distribution.

So this helps us simulate many such numbers within a short amount of time and of course we have the power to simulate the way we want to simulate. And finally plot it so that we can take a look at how such a distribution is. So let me quickly run this program.

(Refer Slide Time: 33:25)

And the figure that you are seeing here on your screen right now is that the red is for the x variable while the green is for the y variable and since I gave the same average and standard deviation. So for this measurement I use the average value of x to be 0, average value of y to be 0 and the standard deviation of x and y to be same = 1 unit. So therefore what ends up happening is that you get a distribution that out of which both onvelope each other meaning that x and y onvelope each other.

So just to give an example of what happens when you change the standard deviation instead of having a standard deviation of 1 what happens if you have a standard deviation of 0.5. So let us simulate it again.

(Refer Slide Time: 34:08)

So what ends up happening here is that you see that for the x variable where we reduce the standard deviation by half you are able to see that the width of this distribution becomes smaller and not just that you are also able to see that the intensity or the number that comes on the y axis is higher for the red than for the blue.

So basically this makes sense since the area under the curve for a Gaussian is constant for these 2. Therefore what is going to happen if you are going to reduce the width the height has to increase okay. Now that you have seen it why do not we exemplify this fact further? I am going to reduce the standard deviation of x from 0.5 to 0.1. So basically we started with 1. Now we are 1 10th standard deviation.

(Refer Slide Time: 34:53)

Let us quickly simulate what we are able to realize is that for the red which is the background the width is much lesser than that of the green. So basically here you can say the green measurement that is y has more uncertainty associated than with the x measurement. So let us also simulate a few small changes that come up. Let us say we switch the average I am switching back the standard deviation of both to 1. Let us say I am changing the average of x to 1 while y remains at 0.

(Refer Slide Time: 35:22)

So now what you end up seeing is that your 2 distributions that come up where the peak value for the green is at 0 well that of red is at 1. So what ends up happening here you still have some values that are on top of each other. Basically there is an overlap between these 2 distribution this indicate that these 2 variables are not very different from each other. You cannot say they are same like in the previous example the same average and standard deviation. But one is able to understand the fact there is some overlap meaning that these 2 variables are not different from one another. On the other hand let us reduce the standard deviation of x back to 0.1.

(Refer Slide Time: 36:01)

Let us see how the distributions look in this case as well you are able to realize all the values of x fall within the region of y. This just makes that the fact that x is determined with the lesser amount of uncertainty than y itself. And therefore once again these values are no different. On the other hand let us say we reduce the uncertainty of y also to the same amount.

(Refer Slide Time: 36:24)

You are going to have distributions that completely fall away from one another. this indicates the fact that these 2 values do not agree with each other and you are not you cannot compare these 2 values. Basically these 2 are 2 different numbers that you obtain which is why many times as an analytical chemist when you view out a number and number without a standard deviation makes very little sense. So to finish off let us once again simulate it with the same kind of variables that we had before 0 for average and 1 for standard deviation.

(Refer Slide Time: 37:00)

Here you are able to see the fact that the distributions overlap with each other with the average and the standard deviation being the same. And the next example that we saw where the standard deviation was lesser however the mean was the same.

(Refer Slide Time: 37:17)

And you still end up saying that the mean value remains same for these 2 distributions.

(Refer Slide Time: 37:28)

And another case where you have little overlap with but the average values are quite different from each other where these distributions start to be different from one another.