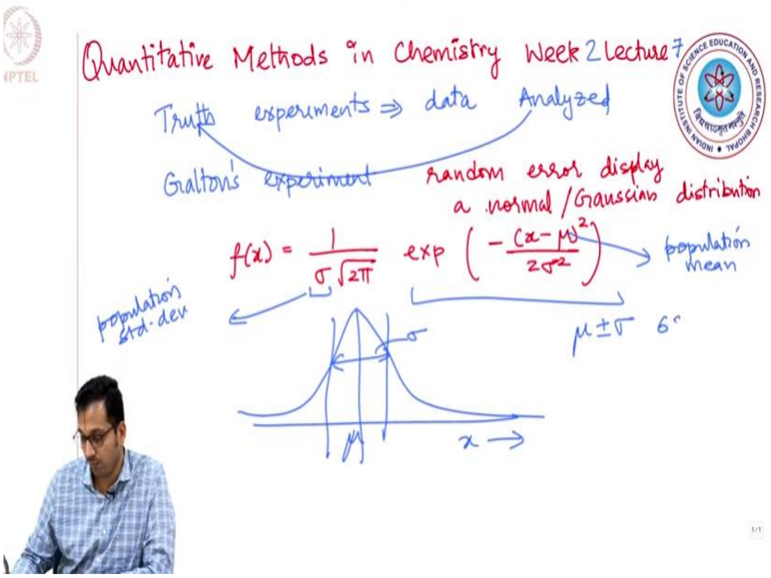
Quantitative Methods in Chemistry

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

Lecture – 07
Using a Spreadsheet towards Basic Statistical Analysis

(Refer Slide Time: 00:29)

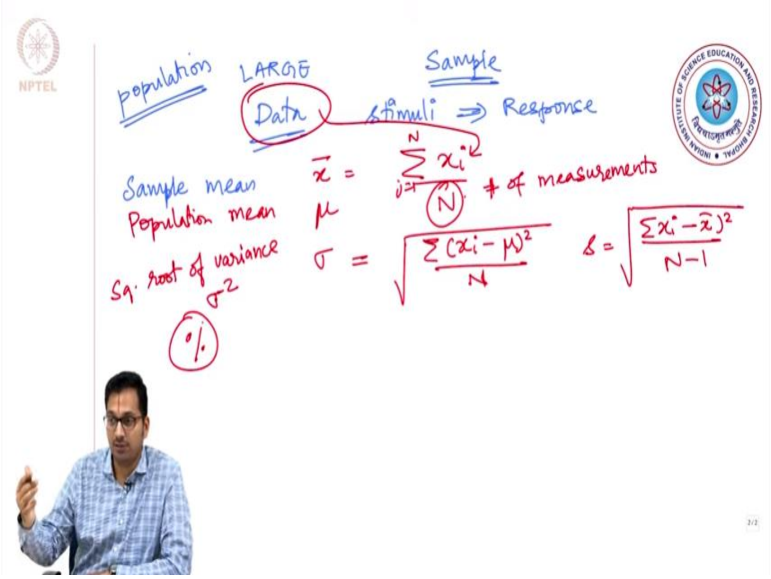

Quantitative Methods in Chemistry Week 2 Lecture 7
Truth experiments ⇒ data Analyzed
Galton's experiment
random error display
a normal / Gaussian distribution
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
population mean
population std. dev
$\mu \pm \sigma$ 6σ
σ
μ
x →
1/1

Welcome to the next lecture in the series of the course quantitative methods and chemistry. So let us quickly recap what we saw in the previous class we defined a few terms. So we wanted to seek the truth in way of experiments and experiments fetches data and we are trying to understand how data can be analyzed to of course go back to understand the truth in this aspect we defined a few terms we try to understand what is the what was the Galton's experiment.

We understood the fact that random errors have or display a normal or a Gaussian distribution the Gaussian distribution. Let us say a function f of x is defined as 1/sigma times square root of 2 pi e power - x - mu the whole square divided by 2 sigma square where sigma is the population standard deviation mu is the population mean and basically this is a normalization that a constant that goes and this is the basic distribution.

So what this means is that when you are making a measurement you have a probability of getting your data point within this distribution and that is determined or rather that is kind of indicated by the standard deviation that comes in and we learnt that measurements within 1 standard deviation correspond to about 68% of data within 2 standard deviations correspond to 95% of data and within 3 standard deviations correspond to 99.7% of data. So this gives you an understanding that if you do get numbers where can it fall and what does it mean okay.

(Refer Slide Time: 02:58)



population LARGE Data Sample

stimuli ⇒ Response

Sample mean $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ # of measurements

Population mean μ

Sq. root of variance σ^2 $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$ $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$

%

So we were talking about population versus sample, population represents a large set but not always we have the luxury of getting the data of the entire population. So what we end up doing is to sample a subset of this population. So therefore we assume that the sample does represent what the average population is also representing. Of course, this is an assumption but not always beyond the luxury to perform infinite experiments.

So therefore the sample is what we have to use in most of our employing most of our experiments. So now that we have understood it so basically let us say you are acquiring data in terms of the fact that when you set up an experiment you have a stimuli meaning that you perturb the system and then you get a response and the response is what is measured in terms of data. It could be spectroscopic technique, it could be a concentration measurement it could be anything of that sort okay.

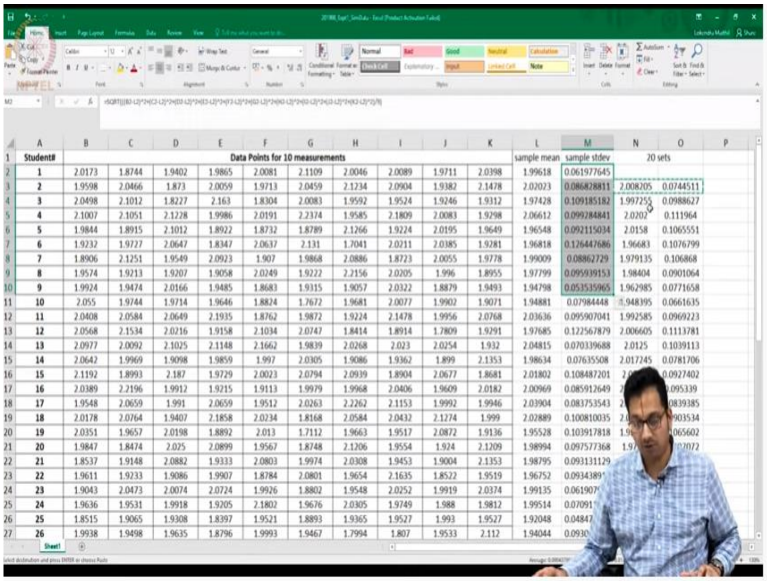
So now one may have this data we defined a few terms here we define what is the sample mean the sample mean was defined as x bar = summation/xi divided by n with i from 1 to n where xi is the data and n is the total number of measurements okay and the only difference between sample mean and population mean is in the fact that n is large for population mean while n is small for the sample mean.

Of course population mean goes with the formula mu which we just saw a moment earlier in the Gaussian distribution and the other definition that we laid forth was the standard deviation. The standard deviation is nothing but the square root of variance, variance is Sigma square, square root of variance and this is defined as the square root of deviations of each data point from the population mean to that of the number of measurements.

Of course when it becomes a small data set it becomes population standard deviation. So that would be given as square root of sum of xi - x bar the whole square divided by n-1 okay so these were the definitions that we saw in the previous class. Let us ask a simple question to start with when you are talking about data and we are talking about percentages. If we make a single measurement what does that mean a single measurement just means that your data could fall within any part of this distribution.

And basically the standard deviation gives you how probable will it fall as close as to the mean for that given measurement and remember the mean the population mean is what helps us seek the truth. Let us say the concentration of a given chemical the mean of multiple measurements should be very close or actually is the concentration of the chemical that you are seeking. So now if you are able to do good number of measurements if you are able to do 1000 measurements it becomes a population standard deviation while if you are able to make only 10 such measurements it becomes a sample standard deviation.

(Refer Slide Time: 06:45)



M2 =SQRT(((B2-L2)^2+(C2-L2)^2+(D2-L2)^2+(E2-L2)^2+(F2-L2)^2+(G2-L2)^2+(H2-L2)^2+(I2-L2)^2+(J2-L2)^2+(K2-L2)^2)/9)

Student#	Data Points for 10 measurements										sample mean	sample stdev	20 sets	
1	2.0173	1.8744	1.9402	1.9865	2.0081	2.1109	2.0046	2.0089	1.9711	2.0398	1.99618	0.061977645		
2	1.9598	2.0466	1.873	2.0059	1.9713	2.0459	2.1234	2.0904	1.9382	2.1478	2.02023	0.086828811	2.008205	0.0744511
3	2.0498	2.1012	1.8227	2.163	1.8304	2.0083	1.9592	1.9524	1.9246	1.9312	1.97428	0.109185182	1.997255	0.0988627
4	2.1007	2.1051	2.1228	1.9986	2.0191	2.2374	1.9585	2.1809	2.0083	1.9298	2.06612	0.099284841	2.0202	0.111964
5	1.9844	1.8915	2.1012	1.8922	1.8732	1.8789	2.1266	1.9224	2.0195	1.9649	1.96548	0.092115034	2.0158	0.1065551
6	1.9232	1.9727	2.0647	1.8347	2.0637	2.131	1.7041	2.0211	2.0385	1.9281	1.96818	0.126447686	1.96683	0.1076799
7	1.8906	2.1251	1.9549	2.0923	1.907	1.9868	2.0886	1.8723	2.0055	1.9778	1.99009	0.08862729	1.979135	0.106868
8	1.9574	1.9213	1.9207	1.9058	2.0249	1.9222	2.2156	2.0205	1.996	1.8955	1.97799	0.095939153	1.98404	0.0901064
9	1.9924	1.9474	2.0166	1.9485	1.8683	1.9315	1.9057	2.0322	1.8879	1.9493	1.94798	0.053535965	1.962985	0.0771658
10	2.055	1.9744	1.9714	1.9646	1.8824	1.7672	1.9681	2.0077	1.9902	1.9071	1.94881	0.07984448	1.948395	0.0661635
11	2.0408	2.0584	2.0649	2.1935	1.8762	1.9872	1.9224	2.1478	1.9956	2.0768	2.03636	0.095907041	1.992585	0.0969223
12	2.0568	2.1534	2.0216	1.9158	2.1034	2.0747	1.8414	1.8914	1.7809	1.9291	1.97685	0.122567879	2.006605	0.1113781
13	2.0977	2.0092	2.1025	2.1148	2.1662	1.9839	2.0268	2.023	2.0254	1.932	2.04815	0.070339688	2.0125	0.1039113
14	2.0642	1.9969	1.9098	1.9859	1.997	2.0305	1.9086	1.9362	1.899	2.1353	1.98634	0.07635508	2.017245	0.0781706
15	2.1192	1.8993	2.187	1.9729	2.0023	2.0794	2.0939	1.8904	2.0677	1.8681	2.01802	0.108487201	[illegible]	0.0927402
16	2.0389	2.2196	1.9912	1.9215	1.9113	1.9979	1.9968	2.0406	1.9609	2.0182	2.00969	0.085912649	[illegible]	[illegible]
17	1.9548	2.0659	1.991	2.0659	1.9512	2.0263	2.2262	2.1153	1.9992	1.9946	2.03904	0.083753543	[illegible]	[illegible]
18	2.0178	2.0764	1.9407	2.1858	2.0234	1.8168	2.0584	2.0432	2.1274	1.999	2.02889	0.100810035	[illegible]	[illegible]
19	2.0351	1.9657	2.0198	1.8892	2.013	1.7112	1.9663	1.9517	2.0872	1.9136	1.95528	0.103917818	[illegible]	[illegible]
20	1.9847	1.8474	2.025	2.0899	1.9567	1.8748	2.1206	1.9554	1.924	2.1209	1.98994	0.097577368	[illegible]	[illegible]
21	1.8537	1.9148	2.0882	1.9333	2.0803	1.9974	2.0308	1.9453	1.9004	2.1353	1.98795	0.093131129	[illegible]	[illegible]
22	1.9611	1.9233	1.9086	1.9907	1.8784	2.0801	1.9654	2.1635	1.8522	1.9519	1.96752	[illegible]	[illegible]	[illegible]
23	1.9043	2.0473	2.0074	2.0724	1.9926	1.8802	1.9548	2.0252	1.9919	2.0374	1.99135	[illegible]	[illegible]	[illegible]
24	1.9636	1.9531	1.9918	1.9205	2.1802	1.9676	2.0305	1.9749	1.988	1.9812	1.99514	[illegible]	[illegible]	[illegible]
25	1.8515	1.9065	1.9308	1.8397	1.9521	1.8893	1.9365	1.9527	1.993	1.9527	1.92048	[illegible]	[illegible]	[illegible]
26	1.9938	1.9498	1.9635	1.8796	1.9993	1.9467	1.7994	1.807	1.9533	2.112	1.94044	[illegible]	[illegible]	[illegible]

So this is the example that we saw in the previous class I will already start showing you glimpses of how to use excel in order to make your analysis easier. Here what we see is that we can calculate average and standard deviation, and these are the data points that you have is that you ended up measuring. So the example that we took is aliquoting 2 ml of water from burette in order to calibrate the burette and what you are able to see is small changes that come up instead of taking the burette we can take an example of a micropipette which can aliquot 2m1 out.

So the average basically is nothing but the sum of each one of this it can be done by the sum as in this fashion divided by the total number of measurements in this case 10. So that will be the average and you get a number like 1.9962 and the standard deviation of course will be nothing but square root of difference between a given number minus that of the average that you have determined this is the sample mean that you are having.

So that is the number that is not going to change so that is going to be dollar L dollar 2 indicating that keep that value constant the whole square plus of course you are going to do C2 - the whole square and so on and so forth I am going until K2 - dollar L dollar 2 the whole square divided by n - 1 which will be 9. So the sample mean is given by the sum of all the cells that contain the data that has been recorded divided by the total number of measurements which is 10 in this case well the sample standard deviation is nothing.

But square root the function that is given an excel as SQRT within B2 - L2 where L2 is the sample mean where B2 is that given measurement. So this is nothing but xi – x bar the whole square given as the exponent 2 similarly going until K 2 divided by 9. So now that we have determined this for the 1st cell doing this with a computer for all the other students the other data sets is straightforward you just copy paste the formula you are able to get the average and standard deviation for each of the student independently and what you are able to realize is that within the single standard deviation almost all the averages work out with each other.

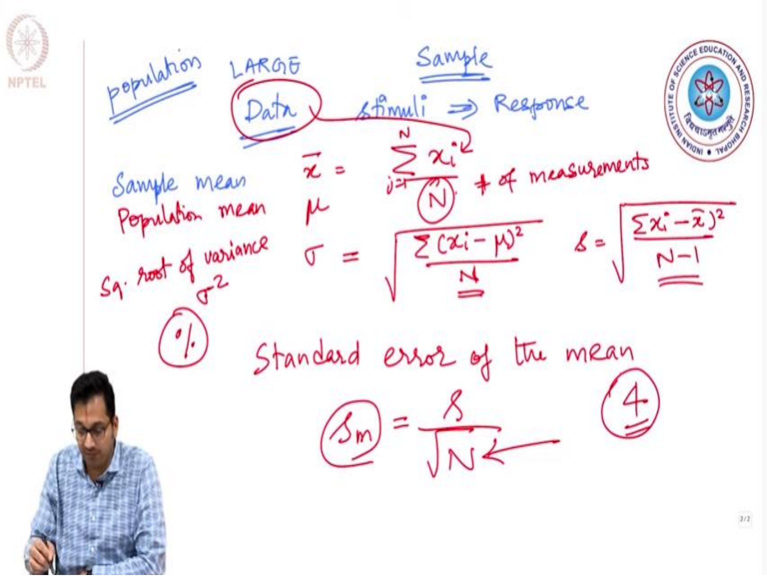
However, there are variations that come in the sample standard deviation just before going ahead and understanding and analyzing this data there are also easier ways of doing this with a software like spreadsheet. So here instead of typing each cell and getting sum as that way you can do a sum and select all of these cells in one shot divided by 10 that is also going to give you the average on the other hand there are functions that have been writ10 in Excel spreadsheets which will help you determine this more easily.

For instance, the average of these numbers which once again is given by average times this will also give something similar. So instead of typing the formula of mean every time you can use some functions that already exist in Excel. Similarly, would be the formula for standard deviation that goes as stdev and then select the number of cells and you get the same number as we have used the formula.

So basically, we have introduced the formula stepwise to you but at the same time how to use spreadsheets to get the answer relatively quickly as well. So now that we have got10 this and let us try to understand instead of using 10 data points in each one of this what happens if you use 20 data points at a given time. So here let us take a look at the average and standard deviation for 20 data sets at a time.

So let me merge the cells to say 20 sets at a time so here I am going to say average and then select these 20 cells and then similarly standard deviation and these 20 cells. So what you are able to realize is that the standard deviation starts to represent more data sets in the average still has not changed much. So let us now copy pasted this once again where it uses 20 data sets at a given time to see how this varies one is able to once again appreciate the fact okay the standard deviations once again are very similar right. So if you keep on changing your 30 days' data sets 40 data sets what happens is the question.

(Refer Slide Time: 11:50)



population LARGE Sample

Data stimuli ⇒ Response

Sample mean $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ # of measurements

Population mean μ

Sq. root of variance σ^2 $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$ $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$

%

Standard error of the mean

$s_m = \frac{s}{\sqrt{N}}$

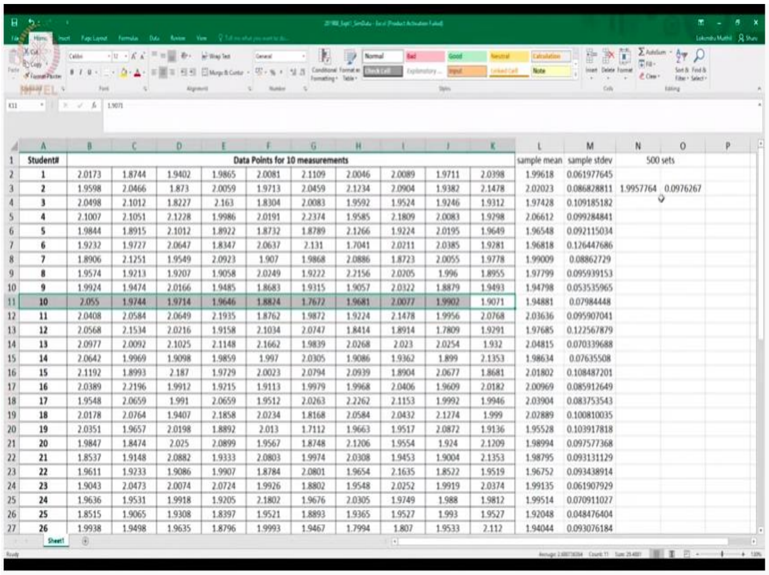
4

So one has to remember that this is called the standard error of the mean meaning if you have s as a standard deviation and the standard error of the mean is given as sm the standard error in this case will go as s divided by square root of n where n is a total number of measurements. So this indicates the fact that as you keep increasing the sample size n your some standard error of the mean keeps reducing.

However, it does not reduce too fast meaning that if you want half your error you need to increase the number of datasets by 4 and if you want to reduce your standard deviation even more significantly you have to go with the square of the precision that you would like to get okay. So now rather than doing this as you are able to see to reduce this is more difficult so people tend to set up experiments such that the standard deviation that comes up is already low, we will be seeing in the next week how this can be done.

So basically this helps you understand how experiments are set and one can also understand that the more number of repeats help you get a better estimate or the reality that is closer to the truth.

(Refer Slide Time: 13:06)



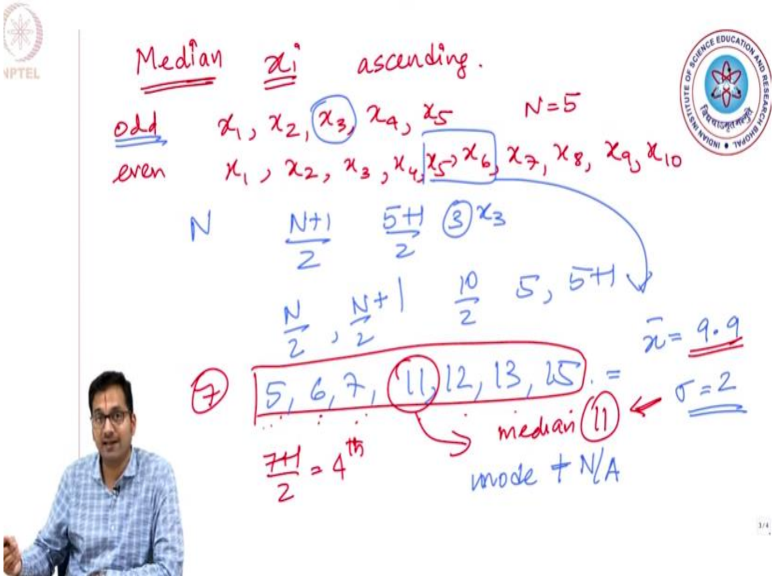
K11 1.9071

Student#	Data Points for 10 measurements										sample mean	sample stdev	500 sets	
1	2.0173	1.8744	1.9402	1.9865	2.0081	2.1109	2.0046	2.0089	1.9711	2.0398	1.99618	0.061977645		
2	1.9598	2.0466	1.873	2.0059	1.9713	2.0459	2.1234	2.0904	1.9382	2.1478	2.02023	0.086828811	1.9957764	0.0976267
3	2.0498	2.1012	1.8227	2.163	1.8304	2.0083	1.9592	1.9524	1.9246	1.9312	1.97428	0.109185182		
4	2.1007	2.1051	2.1228	1.9986	2.0191	2.2374	1.9585	2.1809	2.0083	1.9298	2.06612	0.099284841		
5	1.9844	1.8915	2.1012	1.8922	1.8732	1.8789	2.1266	1.9224	2.0195	1.9649	1.96548	0.092115034		
6	1.9232	1.9727	2.0647	1.8347	2.0637	2.131	1.7041	2.0211	2.0385	1.9281	1.96818	0.126447686		
7	1.8906	2.1251	1.9549	2.0923	1.907	1.9868	2.0886	1.8723	2.0055	1.9778	1.99009	0.08862729		
8	1.9574	1.9213	1.9207	1.9058	2.0249	1.9222	2.2156	2.0205	1.996	1.8955	1.97799	0.095939153		
9	1.9924	1.9474	2.0166	1.9485	1.8683	1.9315	1.9057	2.0322	1.8879	1.9493	1.94798	0.053535965		
10	2.055	1.9744	1.9714	1.9646	1.8824	1.7672	1.9681	2.0077	1.9902	1.9071	1.94881	0.07984448		
11	2.0408	2.0584	2.0649	2.1935	1.8762	1.9872	1.9224	2.1478	1.9956	2.0768	2.03636	0.095907041		
12	2.0568	2.1534	2.0216	1.9158	2.1034	2.0747	1.8414	1.8914	1.7809	1.9291	1.97685	0.122567879		
13	2.0977	2.0092	2.1025	2.1148	2.1662	1.9839	2.0268	2.023	2.0254	1.932	2.04815	0.070339688		
14	2.0642	1.9969	1.9098	1.9859	1.997	2.0305	1.9086	1.9362	1.899	2.1353	1.98634	0.07635508		
15	2.1192	1.8993	2.187	1.9729	2.0023	2.0794	2.0939	1.8904	2.0677	1.8681	2.01802	0.108487201		
16	2.0389	2.2196	1.9912	1.9215	1.9113	1.9979	1.9968	2.0406	1.9609	2.0182	2.00969	0.085912649		
17	1.9548	2.0659	1.991	2.0659	1.9512	2.0263	2.2262	2.1153	1.9992	1.9946	2.03904	0.083753543		
18	2.0178	2.0764	1.9407	2.1858	2.0234	1.8168	2.0584	2.0432	2.1274	1.999	2.02889	0.100810035		
19	2.0351	1.9657	2.0198	1.8892	2.013	1.7112	1.9663	1.9517	2.0872	1.9136	1.95528	0.103917818		
20	1.9847	1.8474	2.025	2.0899	1.9567	1.8748	2.1206	1.9554	1.924	2.1209	1.98994	0.097577368		
21	1.8537	1.9148	2.0882	1.9333	2.0803	1.9974	2.0308	1.9453	1.9004	2.1353	1.98795	0.093131129		
22	1.9611	1.9233	1.9086	1.9907	1.8784	2.0801	1.9654	2.1635	1.8522	1.9519	1.96752	0.093438914		
23	1.9043	2.0473	2.0074	2.0724	1.9926	1.8802	1.9548	2.0252	1.9919	2.0374	1.99135	0.061907929		
24	1.9636	1.9531	1.9918	1.9205	2.1802	1.9676	2.0305	1.9749	1.988	1.9812	1.99514	0.070911027		
25	1.8515	1.9065	1.9308	1.8397	1.9521	1.8893	1.9365	1.9527	1.993	1.9527	1.92048	0.048476404		
26	1.9938	1.9498	1.9635	1.8796	1.9993	1.9467	1.7994	1.807	1.9533	2.112	1.94044	0.093076184		

Sheet1

So instead of taking the 10 data sets at a time we have 50 students in this example and each student has 10 measurements so why do not we take all the 500 data points at the same time so Im going to say average of all these cells and standard deviation of all these cells and what you are able to realize this comes up to be 0.1 okay. So you have subtle variations that come when you have less number of data sets but as you start using more and more number of data sets you tend to get proper representative of the population from the sample itself okay now that we have understood this. Let us try to define a few more terms.

(Refer Slide Time: 14:05)



The next term that we will be introducing to define your data is called the median. Median is defined by the fact that let us say you have a set of numbers xi if you arrange them in the ascending order meaning that from low to high what is the value of the number that comes up in the middle. For instance, let us say you have numbers x1, x2, x3, x4 and x5 and these are your 5 data sets.

So here capital N is set to 5 and you have been able to make 5 such measurements here the median will be the number that comes in the middle. So now let us say we have another data set that goes as x6, x7, x8 and x9 and x10. So in this case we saw an odd number of measurements let us say you have an even number of measurements in this case what ends up happening is that you calculate the median by taking the middle 2 numbers and averaging them.

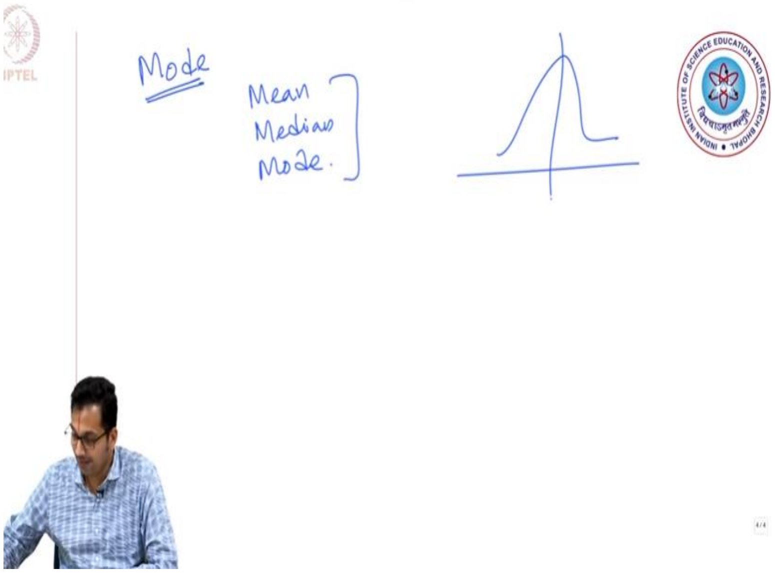
So basically median in the odd case scenario is that let us say you have n measurements that are odd you tend to go n + 1/2 to determine the median. So in this case you had 5 measurements so 5 +1/2 is 3. So therefore you took x3 to assess what is the median if you have even number of measurements then you take the average of the data sets that you got for n/2 and n + n/2+1. so in this case you are going to have what is n/2 your 10, 10/2 is 5. So you take 5 + 5 + 1 so you take the 5th and the 6th point to get what is your median.

Why is this important not always you are going to have a case where all numbers are equally represented? Let us take some examples where you 10d to have a few measurements done in this way 5, 6, 7, 11, 12, 13 and 15 so when you make these measurements, and these are already ordered according in an ascending order. Let us quickly determine what is its average the average for this case is going to be given by close to 9.9 while the median that we will end up measuring, so you have 1,2,3,4,5,6,7 measurements.

So basically it is an odd number, so you go 7 + 1/2 which is the 4th element so 1,2,3 and 4. So you are getting a median as 11 this matters a lot largely because maybe you are looking at the age of students within a given room or when in a crush facility in months. So what is going to end up happening is that you have numbers that go as integer steps, but you get an average that is about 10.

But you are able to see the median that is falling away from the mean that indicates the fact that the middle of the distribution actually does not fall in a mean and that is quite amount of variability that comes of course we have already defined that as the standard deviation that goes. So I leave it as an exercise for you to find the standard deviation, but my guess would say the standard deviation would fall somewhere at 2.

(Refer Slide Time: 17:32)

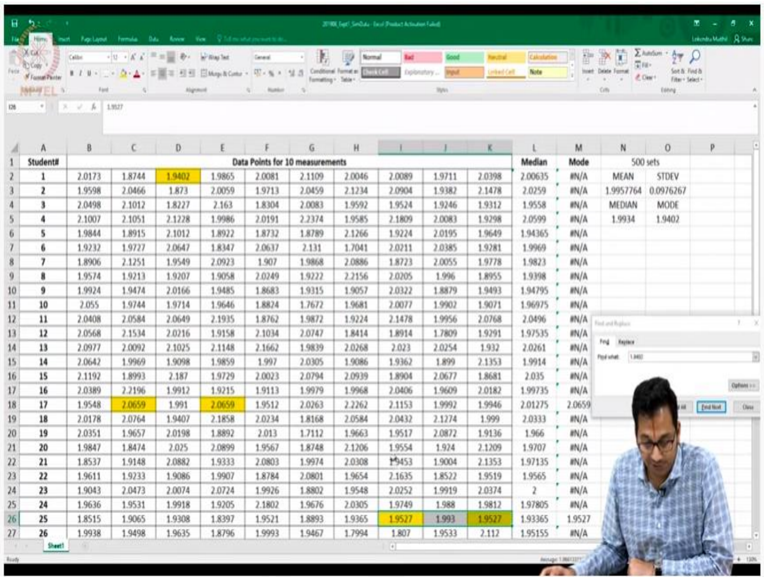


Okay so now that we have seen what is the definition of the median let us take a look at the next definition so in this case, we will be defining what is mode, mode is nothing but an element that is repeated most often in the sequence of data that you are measured. So as in the previous example if we take a close look you realize that none of the measurements have been repeated in 5, 6, 7, 11, 12, 13 and 15 no measurements are repeated.

So in this case mode does not exist its not at applicable mode does not exist okay. But on the other hand what this helps you understand is like the frequency plots that we were seeing yesterday if you are able to find where is the mode and the mode falls very close to the average that indicates which population is skewing the average or which set of sample is skewing the average.

So generally when you are collecting data people tend to report mean, median and mode if all the three agree with each other meaning that the mean mode and median are all close to each other this indicates more or less a proper uniform Gaussian distribution that comes up.

(Refer Slide Time: 18:50)



Student#	Data Points for 10 measurements										Median	Mode	500 sets	
1	2.0173	1.8744	1.9402	1.9865	2.0081	2.1109	2.0046	2.0089	1.9711	2.0398	2.00635	#N/A	MEAN	STDEV
2	1.9598	2.0466	1.873	2.0059	1.9713	2.0459	2.1234	2.0904	1.9382	2.1478	2.0259	#N/A	1.9957764	0.0976267
3	2.0498	2.1012	1.8227	2.163	1.8304	2.0083	1.9592	1.9524	1.9246	1.9312	1.9558	#N/A	MEDIAN	MODE
4	2.1007	2.1051	2.1228	1.9986	2.0191	2.2374	1.9585	2.1809	2.0083	1.9298	2.0599	#N/A	1.9934	1.9402
5	1.9844	1.8915	2.1012	1.8922	1.8732	1.8789	2.1266	1.9224	2.0195	1.9649	1.94365	#N/A		
6	1.9232	1.9727	2.0647	1.8347	2.0637	2.131	1.7041	2.0211	2.0385	1.9281	1.9969	#N/A		
7	1.8906	2.1251	1.9549	2.0923	1.907	1.9868	2.0886	1.8723	2.0055	1.9778	1.9823	#N/A		
8	1.9574	1.9213	1.9207	1.9058	2.0249	1.9222	2.2156	2.0205	1.996	1.8955	1.9398	#N/A		
9	1.9924	1.9474	2.0166	1.9485	1.8683	1.9315	1.9057	2.0322	1.8879	1.9493	1.94795	#N/A		
10	2.055	1.9744	1.9714	1.9646	1.8824	1.7672	1.9681	2.0077	1.9902	1.9071	1.96975	#N/A		
11	2.0408	2.0584	2.0649	2.1935	1.8762	1.9872	1.9224	2.1478	1.9956	2.0768	2.0496	#N/A		
12	2.0568	2.1534	2.0216	1.9158	2.1034	2.0747	1.8414	1.8914	1.7809	1.9291	1.97535	#N/A		
13	2.0977	2.0092	2.1025	2.1148	2.1662	1.9839	2.0268	2.023	2.0254	1.932	2.0261	#N/A		
14	2.0642	1.9969	1.9098	1.9859	1.997	2.0305	1.9086	1.9362	1.899	2.1353	1.9914	#N/A		
15	2.1192	1.8993	2.187	1.9729	2.0023	2.0794	2.0939	1.8904	2.0677	1.8681	2.035	#N/A		
16	2.0389	2.2196	1.9912	1.9215	1.9113	1.9979	1.9968	2.0406	1.9609	2.0182	1.99735	#N/A		
17	1.9548	2.0659	1.991	2.0659	1.9512	2.0263	2.2262	2.1153	1.9992	1.9946	2.01275	2.0659		
18	2.0178	2.0764	1.9407	2.1858	2.0234	1.8168	2.0584	2.0432	2.1274	1.999	2.0333	#N/A		
19	2.0351	1.9657	2.0198	1.8892	2.013	1.7112	1.9663	1.9517	2.0872	1.9136	1.966	#N/A		
20	1.9847	1.8474	2.025	2.0899	1.9567	1.8748	2.1206	1.9554	1.924	2.1209	1.9707	#N/A		
21	1.8537	1.9148	2.0882	1.9333	2.0803	1.9974	2.0308	1.9453	1.9004	2.1353	1.97135	#N/A		
22	1.9611	1.9233	1.9086	1.9907	1.8784	2.0801	1.9654	2.1635	1.8522	1.9519	1.9565	#N/A		
23	1.9043	2.0473	2.0074	2.0724	1.9926	1.8802	1.9548	2.0252	1.9919	2.0374	2	#N/A		
24	1.9636	1.9531	1.9918	1.9205	2.1802	1.9676	2.0305	1.9749	1.988	1.9812	1.97805	#N/A		
25	1.8515	1.9065	1.9308	1.8397	1.9521	1.8893	1.9365	1.9527	1.993	1.9527	1.93365	1.9527		
26	1.9938	1.9498	1.9635	1.8796	1.9993	1.9467	1.7994	1.807	1.9533	2.112	1.95155	#N/A		

Okay now that we have seen that why do not we apply this to the example that were seeing from yesterday’s class. So let us know that we have seen what is the average and standard deviation I am going to be removing the unnecessary elements in this. So now let us ask what is the median for this data set. So for instance let us take one at a time so let us take this data set similar to the functions that I introduced in terms of average and standard deviation.

There is also a function written in spreadsheet generally called median which will help you find the median of this data set what you are able to realize is the median and the mean are very close in this example. So let us calculate it for all the 50 students and what you end up seeing okay you do see some variability but that once again falls within the standard deviation which kind of indicates the fact that its quite normal distribution that you end up getting for the data set that you have seen here this is slightly far away but what you are able to realize within 1.9 to 2.1 all the cases fall in terms of the median analysis.

So now let us ask what is the mode, mode once again is which number gets repeated the most within that given distribution. We are once again using the function that exists in the spreadsheet any instance was not applicable but for some datasets you do end up seeing that you do have a more defined in this case the mode is defined as 2.0659 let us see whether it is indeed correct if you play close attention you see 2 cells here that have 2.0659 that is just a mere coincidence because when measurements are being made you do not have to get the exact same measurement multiple times.

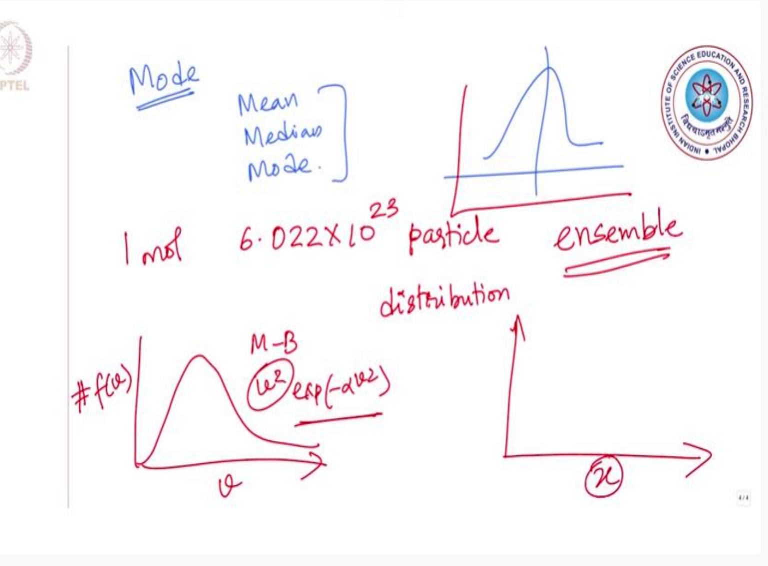
On the other hand, do we have other examples since we do in this case what you are able to see it happens in such a way that these 2 numbers have gotten repeated and you realize for most of the examples you do not have a mode that is properly determined. So now similar to what we did for the 500 data sets let us determine what is the median for all this data set of course what you end up doing is to use all of this and one short please pay attention when you are making such measurements not to choose the wrong columns that end up coming.

That will come with practice and what you are able to see in this case the median nicely matches with the mean once again this indicates that you have a good normal distribution about the mean that you have been able to get and then let us determine what is the mode. So that is going to be equal to this is going to be a very interesting analysis the mode comes us as 1.9402 well. So let us see how many times are we able to spot at least a few cases of such occurrences of one point so you can search.

So select all then you should be able to search the data so let us say find what 1.9402 there you go that is the first occurrence there is the next occurrence. So basically this appears twice and that happens to be interesting. So there are also other times that you get 2 sets of numbers that come in and it tends to pick the lower number indicating that mode in this case does not properly represent the mode that we are trying to look at.

Because there are multiple numbers for instance, we just saw cases where these 2 numbers are also represented 2 times each. So what one thing that you are able to understand is that when you get a data set giving the mean standard deviation median would help you assess what kind of data that you obtained. Now that we have seen how a data is represented properly with mean mode and median and of course standard deviation let us try to ask why are we worrying this much about it.

(Refer Slide Time: 23:19)



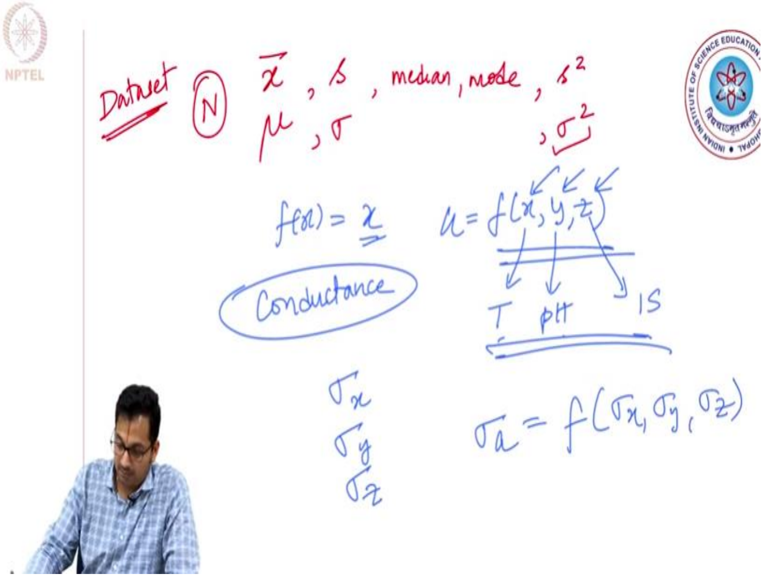
So all these parameters are defined because when you are taking one mode of a substance you are talking about 6.022 into 10 to the power of 23 particles and this means that when you are making a measurement the measurement is never on a single molecule except for a few exquisite experiments which can probe several molecules most often you are going to be looking and an ensemble or a group of molecules.

And you are interrogating all of them together at the same time and of course they themselves have a distribution within them this is not new for chemists who have learnt the kinetic theory of gases where when you plot the number of molecules that possess a velocity v it goes as the Maxwell's Boltzmann distribution which is slightly different from the normal distribution because as it goes from proportional to v square exp - alpha v square here this is the gaussian part but you also have a polynomial part.

So that makes this function and a symmetric function but on the other hand many other parameters that you end up measuring for an ensemble of molecules would probably end up having a normal distribution so basically distributions are what happening is that when you make a certain measurement you are not able to interrogate one molecule at a time but you interrogate the entire population in the given time and the variable that you are trying to measure let us say is x you are going to have the population of x that decides where the average falls most often or not.

So therefore what ends up happening is that these kind of distributions help you assess how far are you from the truth.

(Refer Slide Time: 25:10)


NPTEL
Dataset
N
$\bar{x}$, s, median, mode, s^2
μ, σ, σ^2
$f(x) = x$
$u = f(x, y, z)$
Conductance
T pH IS
σ_x
σ_y
σ_z
$\sigma_u = f(\sigma_x, \sigma_y, \sigma_z)$

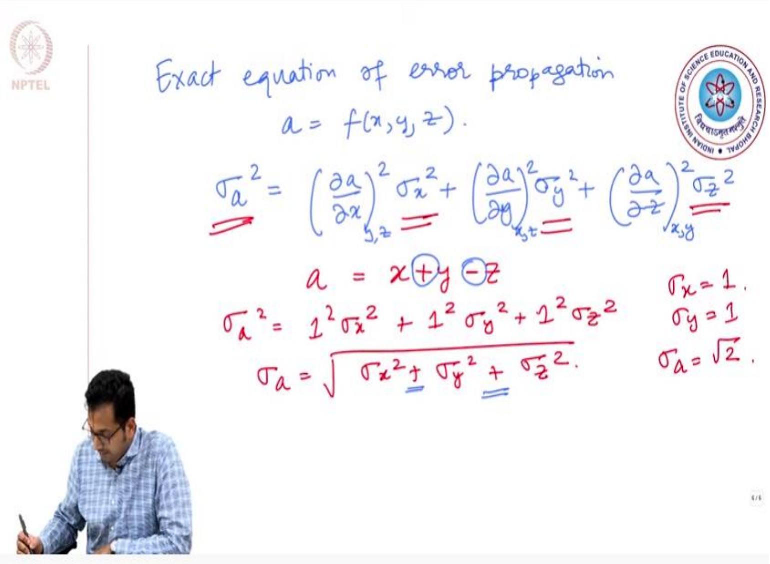
So now we have realized if you have a data set you would have to provide the mean either the sample mean or the population mean depending upon the number of points that you are able to measure and of course the number of points measured n should also be defined when you define the standard deviation once again the sample standard deviation or the population standard deviation and the other parameters such as median, mode.

One parameter that also matters here is the variance which goes as s square for a population for a population measurement then it goes as sigma square. The important point that comes up right now is that most often in science we see parameters that are not a direct dependence you are seldom going to have f of x=x you end up having something like f of x is equal a rather a given parameter is dependent on multiple variables.

Such as f of x, y, z so when this happens you have a variation that comes in the measurement that you are trying to do and that depends on the variation that comes from each of this independent parameters x y and z. So basically what ends up happening if you are able to let us say this could be temperature this could be pH and this could be let us say ionic strength. So what ends up happening if you trying to measure conductance all these variables would end up affecting the final measurement.

So these are all independent variables while conductance is a dependent variable on these independent parameters. So what ends up happening we are able to immediately understand that the uncertainty that will be associated with each of the parameters is going to overall determine the uncertainty of the parameter that you determine. Let us say a is equal to f of x, y, z then the uncertainty with the measurement of a is going to go as a function of all of this.

(Refer Slide Time: 27:24)



Exact equation of error propagation

$$a = f(x, y, z).$$

$$\underline{\underline{\sigma_a^{\,2}}} = \left(\frac{\partial a}{\partial x}\right)^2_{y,z} \underline{\underline{\sigma_x^{\,2}}} + \left(\frac{\partial a}{\partial y}\right)^2_{x,t} \underline{\underline{\sigma_y^{\,2}}} + \left(\frac{\partial a}{\partial z}\right)^2_{x,y} \underline{\underline{\sigma_z^{\,2}}}$$

$$a = x \textcircled{+} y \textcircled{-} z$$

$$\sigma_a^{\,2} = 1^2 \sigma_x^{\,2} + 1^2 \sigma_y^{\,2} + 1^2 \sigma_z^{\,2}$$

$$\sigma_a = \sqrt{\sigma_x^{\,2} \underline{\underline{+}} \sigma_y^{\,2} \underline{\underline{+}} \sigma_z^{\,2}}.$$

$\sigma_x = 1.$

$\sigma_y = 1$

$\sigma_a = \sqrt{2}.$

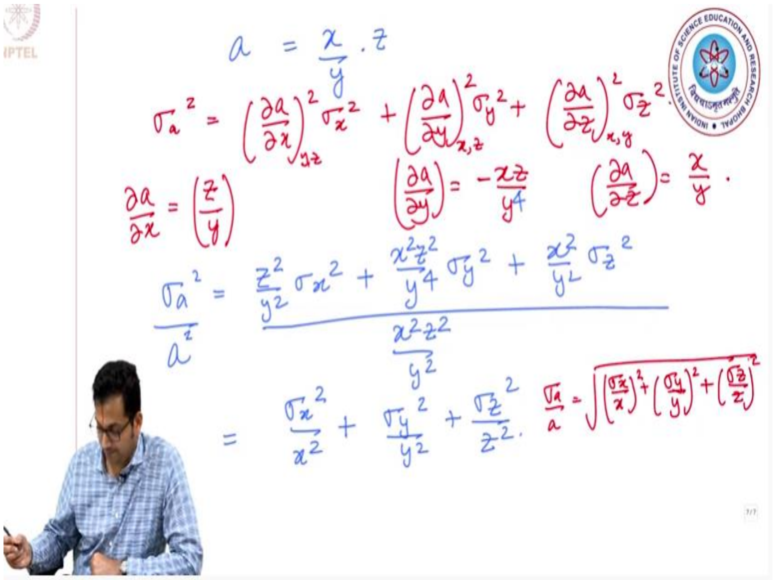
So this goes back to the equation called the exact equation of error or uncertainty propagation so this goes as let us say you have once again a = f of x, y, then sigma a square is going to be given by dou a/dou x the whole square sigma x square + dou a/dou y the whole square sigma y square plus dou a/dou z the whole square sigma z square. So what you are able to realize is for each of the independent variable that is associated with a certain variance the variance of the final measure a will be determined let us take a few examples and take a look at it.

Let us say a is given as x + y- z so in this case why do not we determine what is the standard deviation that is going to what is sigma a going to be. So sigma a square is going to be dou a/dou x. So dou a/dou x gamma bar in this case y and z are kept constant here x and z are kept constant here x and y are kept constant so what will end up happening here is that dou a/dou x is 1. So you are going to have 1 square sigma x square similarly will be the case for this sigma y square +.

So therefore what ends up happening sigma a is going to be given by square root of sigma x sigma x square sigma y square + sigma z square. So what you are able to realize is let us say that we are taking only two measurements and both sigma x and sigma y are one-unit sigma a is going to be given by a square root of 2. So therefore what it means is that although you are adding these 2 the errors do not add up linearly.

This is a very important point for one to understand and at the same time the other take-home message from this example is that irrespective whether you add, or you subtract the standard deviation or the variance that comes up ends up being some of these. So let us take another example.

(Refer Slide Time: 29:53)



$$a = \frac{x}{y} . z$$

$$\sigma_a^2 = \left(\frac{\partial a}{\partial x}\right)^2_{y,z} \sigma_x^2 + \left(\frac{\partial a}{\partial y}\right)^2_{x,z} \sigma_y^2 + \left(\frac{\partial a}{\partial z}\right)^2_{x,y} \sigma_z^2 .$$

$$\frac{\partial a}{\partial x} = \left(\frac{z}{y}\right) \qquad \left(\frac{\partial a}{\partial y}\right) = -\frac{xz}{y^4} \qquad \left(\frac{\partial a}{\partial z}\right) = \frac{x}{y} .$$

$$\frac{\sigma_a^2}{a^2} = \frac{\frac{z^2}{y^2}\sigma_x^2 + \frac{x^2z^2}{y^4}\sigma_y^2 + \frac{x^2}{y^2}\sigma_z^2}{\frac{x^2z^2}{y^2}}$$

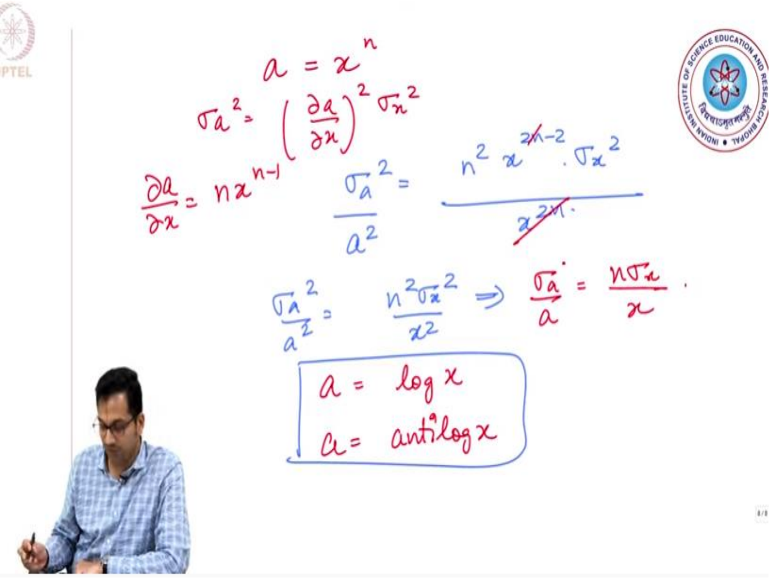
$$= \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2} + \frac{\sigma_z^2}{z^2} . \qquad \frac{\sigma_a}{a} = \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2 + \left(\frac{\sigma_z}{z}\right)^2}$$

In this example let us say a = x divided by y at times z. So once again let us write the formula the exact formula of error propagation sigma a square is going to be given by dou a/dou x the whole square sigma x square + dou a/dou y the whole square sigma y square + dou a/dou z the whole square sigma z square. So now dou a/dou x is going to be given by z of course here y and z are constant x and z are constant x and y are constant.

So therefore this is going to be equal to z/y dou a/dou y is going to be equal to minus xz/y square and dou a/dou z is going to be equal to x/y. Therefore, sigma a square is going to be equal to z square/y square times sigma x square plus because of sum the square of minus xz/y square will be +x square z square/y square y square + x square/y square sigma z square. So let us divide the entire term by a or even a square.

So what will end up happening is divided by x square z square/y square. So immediately one is able to realize when we minimize this it is going to be sigma x square/x square plus sigma y square/y square + sigma z square/z square. So therefore this reduces to sigma a/a = square root of sigma x/x the whole square + sigma y/y the whole square + sigma z/z the whole square alright.

(Refer Slide Time: 32:20)



$$a = x^n$$

$$\sigma_a^2 = \left(\frac{\partial a}{\partial x}\right)^2 \sigma_x^2$$

$$\frac{\partial a}{\partial x} = n x^{n-1}$$

$$\frac{\sigma_a^2}{a^2} = \frac{n^2 x^{\cancel{2n}-2} \cdot \sigma_x^2}{\cancel{x^{2n}}}$$

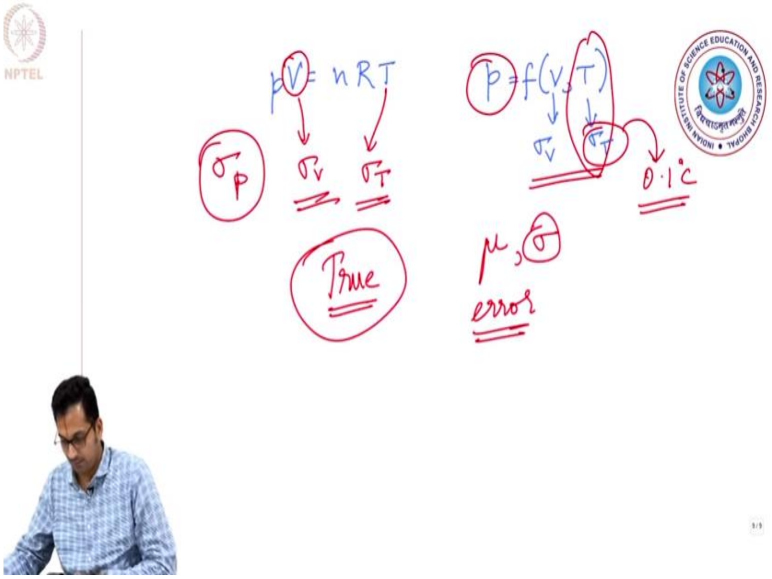
$$\frac{\sigma_a^2}{a^2} = \frac{n^2 \sigma_x^2}{x^2} \Rightarrow \frac{\sigma_a}{a} = \frac{n\sigma_x}{x}.$$

$$\boxed{\begin{array}{l} a = \log x \\ a = \text{antilog}\, x \end{array}}$$

a = x power n where n is a constant. so what is going to happen here sigma a square is going to be given by dou a/dou x the whole square times sigma x square. So dou a/dou x is going to be n into x power n – 1. So this will make sigma a square will be equal to n square x power 2n - 2 times sigma x square. Let us divide the entire term by a square that is going to be equal to x power 2n.

So this reduces we you should be able to cancel this and this so this will result in sigma a square/ a square =n square sigma x square divided by x square. This implies sigma a/a = n sigma x square/x okay sorry n sigma x/x good. I would leave it as an assignment for you to figure out what happens when you want to take a is equal to log of x and let us say a is equal to anti-log of x. Kindly do these two assignments to determine what is sigma a in this case and sigma a in this case as well.

(Refer Slide Time: 33:58)



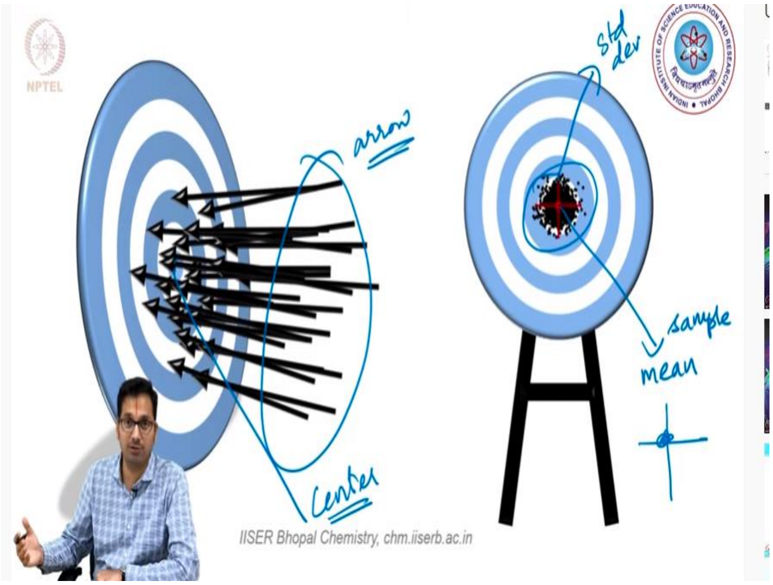
So one is able to understand let us take a simple example so that we try to impress upon why is that we are discussing about this. Let us take the ideal gas equation where pv = nRT to realize that let us say you are trying to determine pressure as a function of volume and temperature and let us say these are the two variables that you have control over, and you have a certain error of measurement that is associated with it.

What do we mean by this is that when you are trying to measure temperature you will be let us say you are using a thermometer in this case and your thermometer has divisions that go between say 0.1 degrees Celsius that means that your measurement cannot be more accurate than the minimum value the least count that you can use for this such a measurement. So what ends up happening by default the pressure is going to be having uncertainties from volume and from temperature.

But the exact equation of error propagation helps you determine and quantify if you know if you are able to quantify sigma V and sigma T you can precisely quantify what is sigma p. So this is an important part of understanding how errors go forward in the following lecture well be understanding how this will all help us assess what is the true value. So this is how we started the lecture we were trying to say we are trying to seek the truth in terms of experiments by collecting data.

While we collect data you get something like a mean and a standard deviation and this standard deviation be for multiple independent variables that you are varying to get your measurement and we are able to see through error propagation how finally for a given variable that you want to measure these errors propagate. Now that we see it, we will try to define what is the true value and what is accuracy versus precision.

(Refer Slide Time: 35:53)



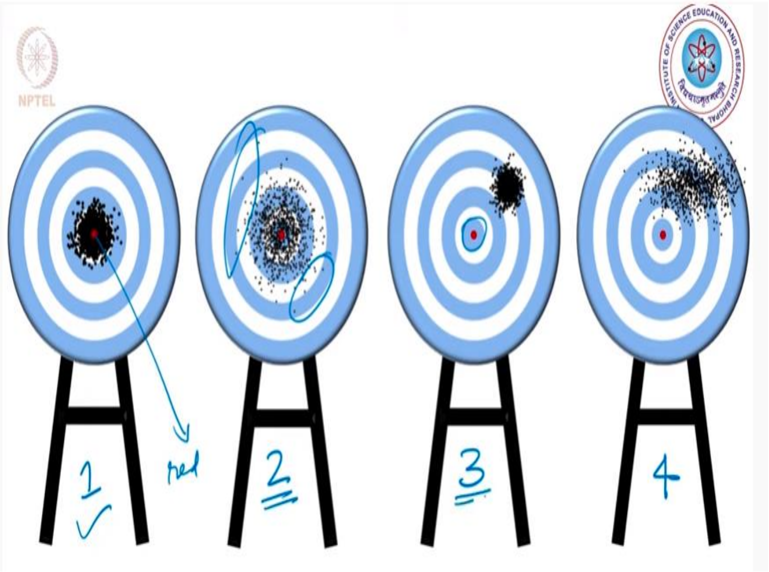
In order to do this, let us take a example for this example let us think that there is an archer who has a bow and then of course he is trying to use the arrows to pierce into the center of this target okay. So what is going to end up happening he is going to first put the arrow here and the bow is going to hit it and we are going to ask let us say the archer is trying to aim how far is his aim good okay.

So this is nothing but the measurement that we are trying to do let us assume that our true value exists at the center of this board and we are the experimentalist we are getting each one of the data points which can be thought about as the arrows that are being pierced onto the board. How close are we getting to the center of the board of course this depends on how good the archer is able to aim and able to hit the center.

So now this each of this arrow points is going to be depicted by the black dots on this board and the red dot indicates what is that mean, the sample mean and of course if you have many measurements it is going to become the population mean what is the sample mean that goes which such an archer aiming in the middle of the board and you have a bar that comes along with the average this depicts the standard deviation.

So basically you are able to understand is that is that if the archer is fraying throughout the board the standard deviation is going to be more but if you are able to take an average that finally works out to be the center of the board you are still happy as the average works out to the true value but you would want improve it because the spread is a little more.

(Refer Slide Time: 37:46)

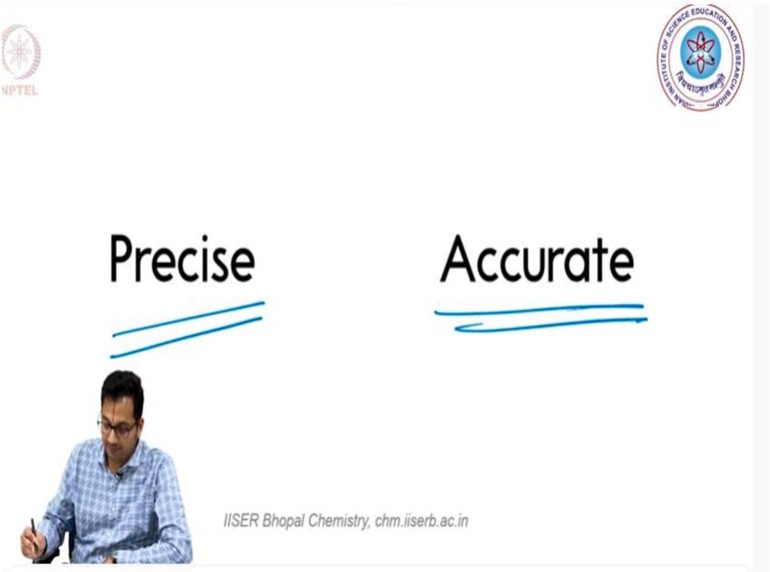


So you see a few examples for such a scenario. So here you are able to see 4 different boards 1 2 3 and 4 and these cases let us take a little while to understand what is going on. In each of the case the true value is represented by the red dot in the center and the scatter of the black dots indicate where the arrow has hit and in this case what you are able to see visually already is that all the dots are close to the center and there are very few dots that go outside the 2nd circle that is in blue.

On the other hand, you are able to see in the 2nd board the dots are still close to the center spread around the center. But there are a few dots that were really farther away so qualitatively one can already say that archer 1 seems to be better than archer 2 largely because the spread is lesser, and both average out to be in the center of the board.

On the other hand, if you see the true values in the center archer 3 unfortunately has been able to get all the points concentrated on one side of the board unfortunately its far away from the center of the board but still all these values are concentrated on one side. On the other hand, archer 4 is the poorest of all because all the points are spread over in all different places and even some of them have gone outside the board and you are able to realize neither you are having all the points being bunched up together and it does its not even spread close to the center.

(Refer Slide Time: 39:12)



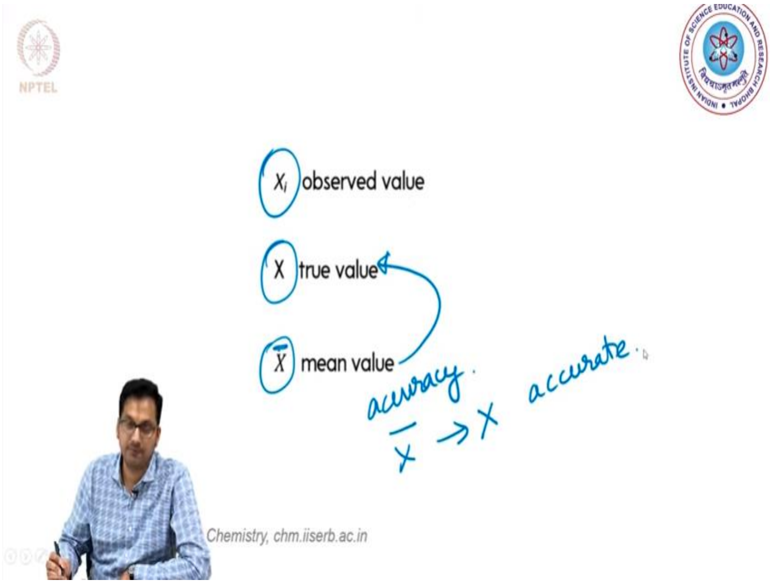
Precise

Accurate

IISER Bhopal Chemistry, chm.iiserb.ac.in

Here is where we are going to be defining the two very important parameters of precision and accuracy.

(Refer Slide Time: 39:20)



X_i observed value

X true value

$\bar{X}$ mean value

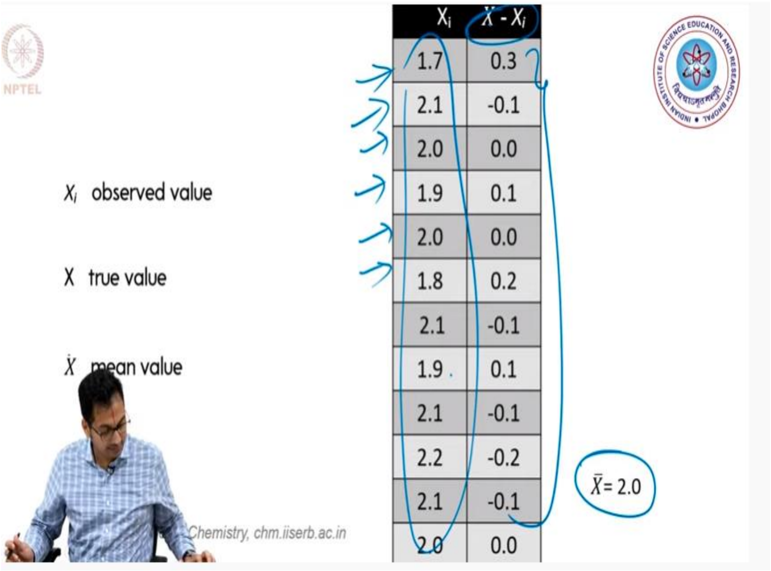
accuracy.

$\bar{x} \rightarrow x$ accurate.

Chemistry, chm.iiserb.ac.in

So what do we mean by these let us say xi is the observed value this is the way we have been defining it so far and x is the true value and x bar is the mean sample mean that we end up getting how close is the mean value to the true value indicates accuracy. The more closer x bar is to x you try to say that measurement is accurate.

(Refer Slide Time: 39:53)



X_i observed value

X true value

$\bar{X}$ mean value

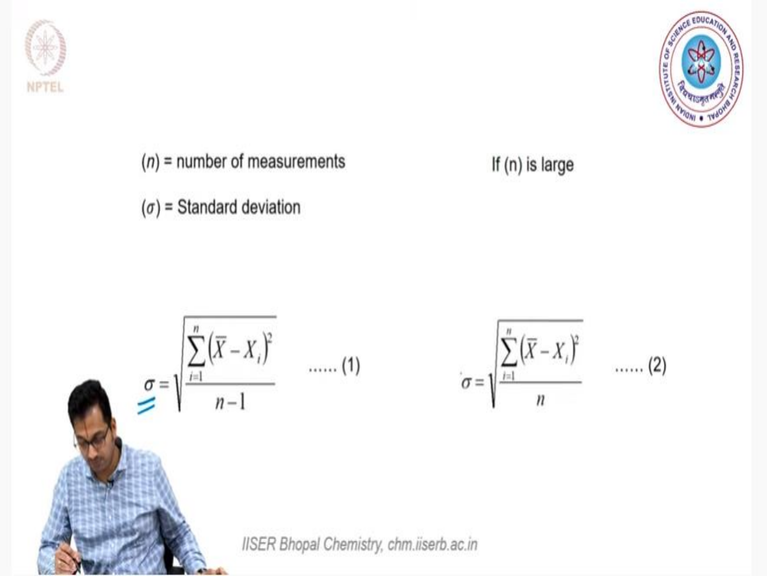
X_i	$\bar{X} - X_i$
1.7	0.3
2.1	-0.1
2.0	0.0
1.9	0.1
2.0	0.0
1.8	0.2
2.1	-0.1
1.9	0.1
2.1	-0.1
2.2	-0.2
2.1	-0.1
2.0	0.0

$\bar{X} = 2.0$

Chemistry, chm.iiserb.ac.in

On the other hand, so let us take one another example so let us say once again you are trying to measure 2 ml out of a burette and these are the measurements that you have done just as the example that we have seen and this gives you an idea how far you are going away from the mean of all the values that have come up or rather from the true value that has come up.

(Refer Slide Time: 40:09)



(n) = number of measurements

(σ) = Standard deviation

If (n) is large

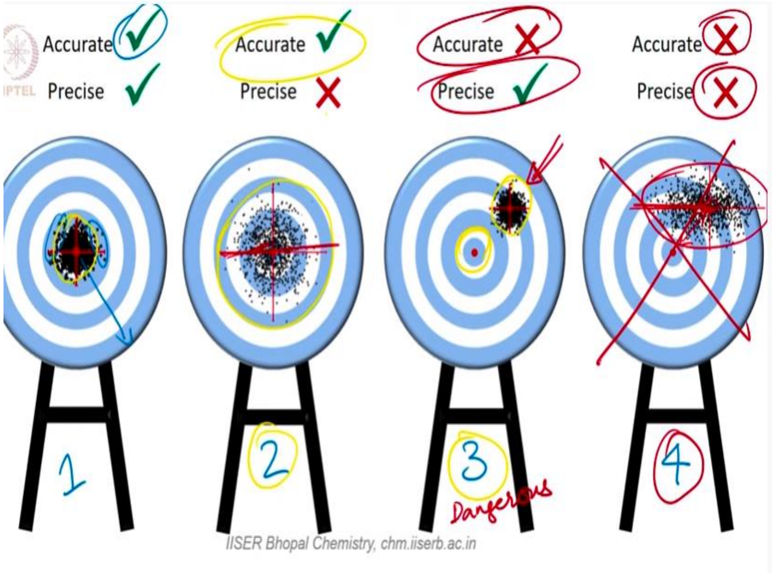
$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(\overline{X}-X_i\right)^2}{n-1}} \quad \text{...... (1)}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(\overline{X}-X_i\right)^2}{n}} \quad \text{...... (2)}$$

IISER Bhopal Chemistry, chm.iiserb.ac.in

And then you realize that can I make an assessment that goes with it. Once again these are the definitions that we have already seen the sample standard deviation will be given such as this and if you have large number of days that becomes the population standard deviation.

(Refer Slide Time: 40:23)



Accurate ✓ Precise ✓

Accurate ✓ Precise ✗

Accurate ✗ Precise ✓

Accurate ✗ Precise ✗

IISER Bhopal Chemistry, chm.iiserb.ac.in

So now let us try to define how accurate are archers 1, 2, 3 and 4 with the examples that we have taken so far. Of course the true value is the center of the board so now what an ends up happening if you take the average of all these data points which I have done, and I am representing by this red dot. You see that it nicely matches with the center of the board then you say the archer is accurate and let us say this is the best of a precision one could get meaning that the archer has not gone beyond the 2nd circle.

If you see carefully almost all the data points lie within the 1st circle and only a few data points are lying outside. Once again this makes total sense because when you are talking about the distribution you are having almost 66 to 68% of the points falling within the 1st standard deviation and 95 points falling within 2 standard deviations and 99.7 points falling within 3 standard deviations.

On the other hand, let us go to the next archer here you are able to see that the points are spread until 1, 2, 3, 4 5 you even see a data point in the 6th circle but nicely the average works out to be in the center. But on the other hand let us determine what is the standard deviation that comes up you are able to realize the standard deviation is much farther while in this case the standard deviation was much lesser, so this says archer 2 is less precise than archer 1.

Of course that could be cases where you cannot improve precision in some cases let us say the archer is working under very cold conditions and trying to hit the arrow in the middle yeah probably then this is probably the best ever precision one could get. But still on an average the archer is still accurate. So now let us go to the next example where the true values in the center of the board that is archer 3 but the average works out somewhere in the center but if you are able to see the standard deviation is much smaller than that of the 2nd archer.

So this is a case where you have poor accuracy but very good precision. This is a very tricky place to be because many times when somebody does measurements and keeps getting a reproducible values and thinks that is the truth this is the scenario that is extremely dangerous because this tends to tell this is the truth but in reality, your truth is very far away from wherever your data points are. So this could be a case we use once again see in the next week why such a case can arise and how an experimentalist can minimize such mistakes to happen.

On the other hand, let us see the fourth case in the 4th case you see that neither the archer is precise not accurate the accuracy and precision are both poor because the average falls away from the true value meaning that you are poor accuracy and the precision is extremely poor even poorer than what you see here because you see the width of this is much higher than the width of this. So basically this is the case where you are inaccurate and imprecise.

So this is an example of how badly an experiment could be set up and therefore neither you get the truth, nor you did the setup properly. So basically this indicates this is the worst ever case one can be and we should strive to make sure we do not end up in that quadrant and of all the archers we would like to be the 1st archer in the case that we are both accurate and precise and there could be some conditions where we cannot avoid and its okay to be 2nd archer where you still get a decent accuracy but poor precision of course we should strive to make the precision better.

But the last two cases are something that we must avoid as it results in poor interpretation of the data and hypotheses which you will be seeing in the forthcoming weeks that will be handled by Dr. Aasheesh Srivastava. Thank you.