Heat Treatment and Surface Hardening - II

Prof. Kallol Mondal
Prof. Sandeep Sangal
Department of Material Science & Engineering
Indian Institute of Technology, Kanpur

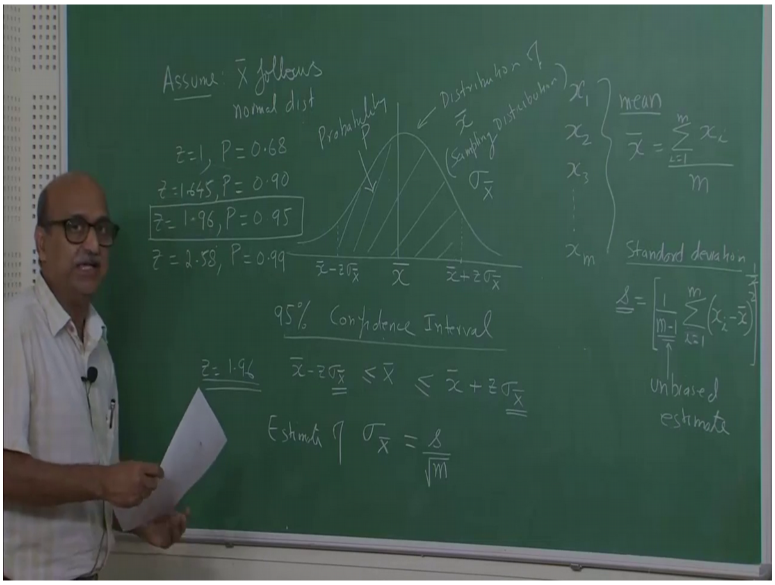
Lecture - 37
Statistical tools for analysis and reporting of obtained data with examples

So, in the last lecture we had completed discussion on the tools stereological tools for the quantification of microstructure. Now in this particular lecture we will look at that if we have made some measurement. We will look at how to get statistical estimates of that measurement, how to get statistical errors associated with the measurement. For example, if I was trying to determine the avrami parameters n and k I would have had to measure volume fraction of the phase transformed for different times and for each time and temperature I would be required to measure using let us say point count technique number of points falling in the phase of interest and from that I am getting the point fraction to get an estimate of the volume fraction.

So, I will make these measurements in several different areas of my sample to get a good statistical estimate in terms of the mean value. That what is the average of all the measurements I have done. And also try to determine what kind of uncertainty statistical uncertainty that is there in my estimates of let us say the mean volume fraction. It could be a problem like a grain size measurement, that after I have given certain kind of heat treatment I have a certain grain size in the material. I want to get an estimate of the grain size using the tools that we have already discussed. And I want to get an uncertainty associated with the with the grain size. So, that I can put some kind of a interval a grain size interval over, which I expect that my true grain size will lie. So, what we will do is first we will have a small theoretical discussion on this and then we will take up one particular problem of volume fraction estimation.

Essentially how to this lecture is essentially how to analyze the data that we get from our measurements. So, let us just first consider a completely theoretical perspective here that any experimental measure that I am making will be subjected to statistical errors statistical fluctuation. For example, I measure let us say volume fraction.

(Refer Slide Time: 03:00)


Assume: $\bar{x}$ follows normal dist
$z = 1, P = 0.68$
$z = 1.645, P = 0.90$
$\boxed{z = 1.96, P = 0.95}$
$z = 2.58, P = 0.99$
Probability P
Distribution of $\bar{x}$ (Sampling Distribution)
$\sigma_{\bar{x}}$
$\bar{x} - z\sigma_{\bar{x}}$
$\bar{x}$
$\bar{x} + z\sigma_{\bar{x}}$
95% Confidence Interval
$z = 1.96$
$\bar{x} - z\sigma_{\bar{x}} \leq \bar{X} \leq \bar{x} + z\sigma_{\bar{x}}$
Estimate of $\sigma_{\bar{x}} = \frac{s}{\sqrt{m}}$
x_1
x_2
x_3
x_m
mean
$\bar{x} = \frac{\sum_{i=1}^{m} x_i}{m}$
Standard deviation
$s = \left[\frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2\right]^{\frac{1}{2}}$
unbiased estimate

So, I in area one I get a volume fraction value of x 1, in area 2 in the sample I get a estimate of x 2. Area 3 I get an estimate of x 3 and so on, and let us say I have taken m such area. So, I will have m such estimates of the volume fraction for that matter. These x 1 x 2 x 3 and x m could have been grain size. They could have been surface area per unit volume.

So, it does not matter what the values are, but these are all experimental measurements. I have made from this set the average value from the experimental measures x 1 x 2 to x m is simply the sum of all of these values divided by m, the total number of measurements. So, I can write x bar the average value as sum of x size from i to m divided by the number of measurements I have made. So, this gives me a point estimate or an average value of the particular measure that I have made on the microstructure whether it is volume fraction or grain size or whatever else. Another thing that we do know that to get some kind of a estimate of the scatter of the values that I have I estimate what is called as a standard deviation. So, this I will call it as the mean value of my experimentally gathered values.

And then from these values I can also obtain what is called as the standard deviation and let me denote it as s, and this from a set of m data points is given by the following relationship. 1 upon m minus 1 summation I equals 1 to m x i minus x bar the mean value squared. So, I am basically what I am doing is I am summing up the squares of the deviations, and then dividing not by m, but I am dividing by m minus 1. And this is an important point to note that dividing by m minus 1 gives me what is called as an unbiased estimate of the standard deviation. And of course, this whole thing has to be taken to the power of half or taking the square root of the entire summation after dividing by m minus 1.

So, this actually is giving me an unbiased estimate of the population standard deviation. So, larger is this value larger is the scatter in my measurement. It perhaps could be because the sample itself has a lot of scatter in the particular measurement I am making for example, volume fraction different regions are having very different volume fraction or if this value is small then it indicates that we have a very uniform sample. Now if I let us say I make one set of measurements I get a particular value of mean and a particular value of standard deviation. Now suppose I were to repeat this measurements again in m other areas in the sample. Will I get the same mean or the standard deviation no I will get somewhat different estimates of the mean value of my let us say the grain size.

I am measuring or the volume fraction I am measuring if I repeat it a third time I will get some other estimate. So In fact, if I made many such sets I will get a distribution of the mean values. This is important to note. So, what kind of a distribution would this mean follow? Well, if we have made sufficiently large number of measurements we can at times assume that the distribution of the means may follow a normal distribution. So, this is distribution of x bar, this is also in statistics called sampling distribution. If let us say it follows a normal distribution. So, let us make an assumption here, assume that x bar follows normal distribution. And let us say the standard deviation of this particular distribution the sampling distribution is denoted as sigma x bar.

Therefore what can I say? If I have some estimate of sigma x bar, and I have an estimate of x bar which is over here. What can I say regarding some kind of an interval over which I expect that the mean value of the population will lie? So, one property of the standard deviation of the normal distribution is that if I take a number z, and let us take a point here which is x bar the mean of this particular distribution plus z times sigma x bar. And take another point on the left of x bar which is mean minus some same number z times the standard deviation, then irrespective of the mean or the standard deviation. One property is that the area in this region is fixed for a given value of z.

And what does this area represent? This area represents the probability that the mean will lie in the range x bar minus z sigma x bar and x bar plus z sigma x bar. So, this denotes the probability P and this P is going to be a function of z, if a chase z my areas would; obviously, be different now just to give you some idea regarding what are these areas. So, for z is equal to 1 P equals 0.68. Which means that mean minus 1 standard deviation the range mean minus 1 standard deviation to mean plus 1 standard deviation the value will lie in this range 68 percent of the time. So, there is a probability of 0.68 that the value will lie in this range.

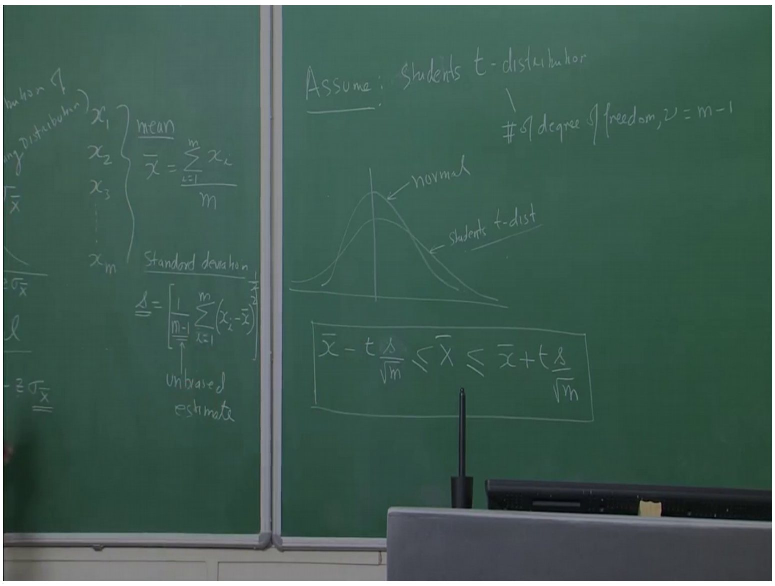
So, this is an interval we have defined for z is equal to 1. Now I can take up some other z values for example, if I take z equals 1.645, then P is equal to 0.90 or 90 percent probability that the value will lie in this range. Z equals 1.96, implies P is equal to 0.95 or 95 percent probability that the value of the mean will lie in mean minus 1.96 times sigma x bar and mean plus 1.96 into sigma x bar. Similarly z equals 2.58 implies a probability of 0.99. Now obviously, you can have many other z values, but very often we are going we use a particular value of z equals 1.96 or 95 percent. So, then if I have from my measurement a mean value and if I can also estimate sigma x bar, then I can obtain what is called as the 95 percent confidence interval for my mean value.

Which means that I expect that x bar or the average measure that I am making the mean of the population capital x bar is expected to lie between the range x bar that I have obtained from a set of measurement minus z sigma x bar, and this is a lower limit the upper limit x bar plus z sigma x bar. And by putting z equals 1.96. I will get what is called this particular interval is called as the 95 percent confidence interval. Or this is a expect this is sometime also referred to as a error bar that one would see. This error bar as you can see if I increase my confidence level, let us say I want to go at 0.99 or 99 percent that I want to be 99 percent sure that the value will lie in a particular range, then I will have to use z is equal to 2.58.

And this range will increase if I go to a lower confidence level like 90 percent this value will shrink. So, there is a trade off on the kind of error bar that I can associate with my measure. Higher the confidence level larger will be the error bar smaller the confidence level smaller would be the error bar. Now next question arises how do I find this sigma x bar well, is statistical analysis shows us an I am not going to give you a proof here, but any statistical textbook would carry this proof, and it is a very straightforward proof that sigma x bar can be estimated. So, estimate of sigma x bar is given by the relationship. Using the standard deviation of my set of measurements s divided by square root of m where m is the number of measurements I have made.

So, that is a pretty straightforward way of determining my confidence interval or the error bar for my measured set of values for some given microstructure. Now I had made a statement saying that I can assume this x bar or the sampling distribution to be a normal distribution provided I have my I have made large number of measurements my m is sufficiently large. But many times my measurements may not be very large. And hence the assumption of a normal distribution may not be valid. In that case we can make an alternate assumption that assumes that sampling distribution follows at what is called as the students t distribution.

(Refer Slide Time: 17:45)


Assume: Students t-distribution
# of degree of freedom, $\nu = m-1$
normal
students t-dist
$\bar{x} - t\frac{s}{\sqrt{m}} \leq \bar{X} \leq \bar{x} + t\frac{s}{\sqrt{m}}$
x_1
x_2
x_3
x_m
mean
$\bar{x} = \frac{\sum_{i=1}^{m} x_i}{m}$
Standard deviation
$s = \left[\frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2\right]^{\frac{1}{2}}$
unbiased estimate
Distribution
$\sigma_{\bar{x}}$

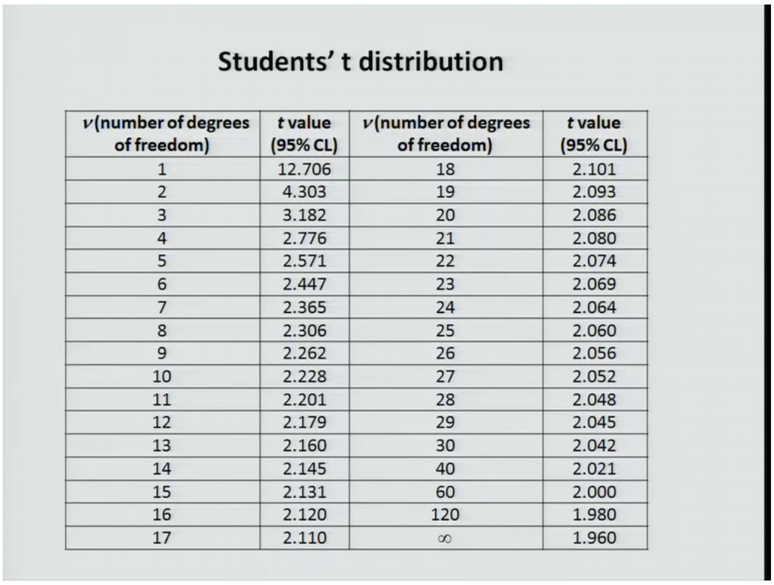
Now the students t distribution, the students t distribution is also very similar to the normal distribution it also has a bell shaped curve, but it is much more spread out.

So, for example, if I want to show the normal distribution in the students t distribution on the same graph, this is the kind of So, this is normal for a given mean and standard deviation. And this is the students t distribution. But as m the number of data points that I have I keep increasing the students t distribution becomes narrower and narrower and for very large m it starts to coincide with the normal distribution and theoretically when m tends to infinity the students t distribution becomes the normal distribution. In that case instead the only difference in finding out the confidence interval would be that instead of using a value of z which is coming from the normal distribution, we have to replace this z with the t value coming from the student's t distribution.

And therefore, in this case the confidence interval would become x bar minus t t times s square root m less than equal to the s mean of my population times x bar plus t times s square root m. Now only thing in this relationship is that the z has been replaced by t. How do I find the t values? Well, for the t distribution there are tables from which we can determine the t distribution and the t distribution as I have already said varies with the number of values that I have measured.

And In fact, t distribution varies with what is called as the number of degrees of freedom nu, which in this kind of problems the number of degrees of freedom nu is simply equal to m minus 1. So, let us look at a table of t values for let us say if I want to have a confidence interval of 95 percent. So, let us see this table here is a table of students t distribution.

(Refer Slide Time: 21:31)



Students' t distribution

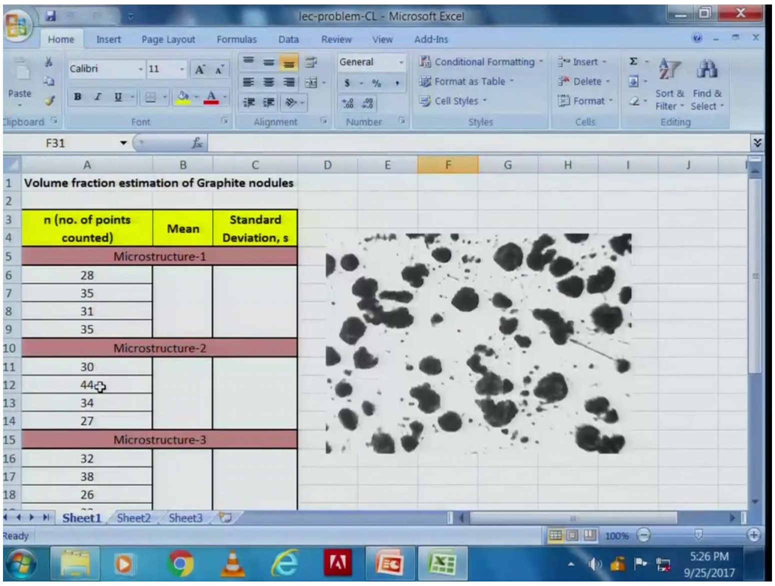
ν(number of degrees of freedom)	*t* value (95% CL)	ν(number of degrees of freedom)	*t* value (95% CL)
1	12.706	18	2.101
2	4.303	19	2.093
3	3.182	20	2.086
4	2.776	21	2.080
5	2.571	22	2.074
6	2.447	23	2.069
7	2.365	24	2.064
8	2.306	25	2.060
9	2.262	26	2.056
10	2.228	27	2.052
11	2.201	28	2.048
12	2.179	29	2.045
13	2.160	30	2.042
14	2.145	40	2.021
15	2.131	60	2.000
16	2.120	120	1.980
17	2.110	∞	1.960

As you can see this is the number of degrees of freedom space this gives you the corresponding t value at 95 percent confidence interval and so on, it goes from one onwards up to 120 and then finally, for very large or infinite value it the t should correspond to the z value for the normal distribution which as you know is 1.96.

So, for example, let us say I have measured 10 different areas of volume fraction, which means the number of degrees of freedom would be 10 minus 1 9. So, I look up on the left on this first column 9 as a number of degrees of freedom and the t values 2.262, if I had let us say 30 as the number of measurements I have made. Then the degrees of freedom would become 29 and the t value would become 2.045. If I had let us say degrees of freedom as 120; that means, I have measured 121 set of values I will have 1.98 as a t value which is now becoming closer and closer to the z value for the normal distribution for 95 percent confidence level.

So, this once I know this I know the t value I can again calculate the confidence interval. So, what we will do now is take up a simple set of experimental data and actually do an actual calculation.

(Refer Slide Time: 23:17)



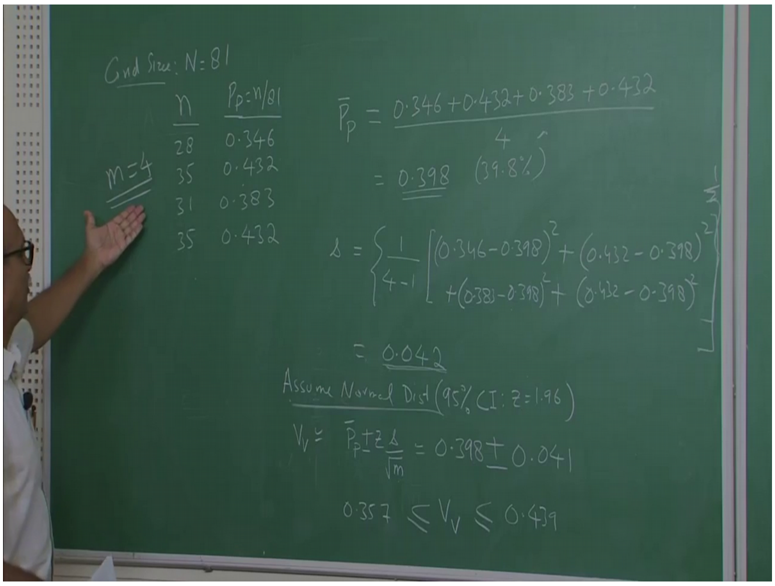
Volume fraction estimation of Graphite nodules

n (no. of points counted)	Mean	Standard Deviation, s
Microstructure-1		
28		
35		
31		
35		
Microstructure-2		
30		
44		
34		
27		
Microstructure-3		
32		
38		
26		

So, let us look at this. Here what we have is we want to estimate the volume fraction of graphite nodules by the point count method, and these measurements have been made on different regions of the microstructure as well as it different microstructures themselves. So, there are 5 different microstructures obtained. And each microstructure a grid of points were placed at different parts and you are only seeing a small part of one of the microstructures that I put here. So, in 4 different areas of a given micrograph the numbers of points intersecting the graphite nodules have been put in under this heading of the first micrograph or the microstructure one.

So, I have got 28 35 31 35. Microstructure 2 I have 344 34 27 and so on. Total of 20; that means, I have actually looked at 20 different areas in the structure. Now I can for purposes of demonstrating the estimate of confidence interval or the error bar. I will just take up the data obtained just 4 values. So, my n is equal to 4 in this case for microstructure 1, and we will follow it up and calculate the result out of that. So, let us consider the data obtained over here for microstructure one and I will just go to the board and do some calculations to show you how to get the confidence interval. So, let us look at this first set of 4 values.

(Refer Slide Time: 25:09)



Grid Size: N=81

$m=4$

n	$P_p = n/81$
28	0.346
35	0.432
31	0.383
35	0.432

$$\bar{P}_p = \frac{0.346+0.432+0.383+0.432}{4} = 0.398 \quad (39.8\%)$$

$$s = \left\{\frac{1}{4-1}\left[(0.346-0.398)^2+(0.432-0.398)^2+(0.383-0.398)^2+(0.432-0.398)^2\right]\right\}^{1/2}$$

$$= 0.042$$

Assume Normal Dist (95% CI: z=1.96)

$$V_V \simeq \bar{P}_p \pm z\frac{s}{\sqrt{m}} = 0.398 \pm 0.041$$

$$0.357 \leq V_V \leq 0.439$$

I have measured n the number of points falling inside the graphite nodules when I place the grid, 28 point sub intersected 35 point sub intersected 31 and 35.

So, 4 measurements I have made if 4 different areas of micrograph 1. And the grid size how many points were there in the grid I will denote that as capital n and that was 81, hence a point fraction from each of these set of data points would be simply equal to n divided by 81. So, 28 divided by 81 would give me 0.346 of 34.6 percent is the estimate of the fraction. 35 divided by 81 would be 0.432, 31 divided by 81 would be 0.383 and finally, 35 well we already have one figure like that. So, it is the same 0.432. Now to get first of all the mean value of the point fraction.

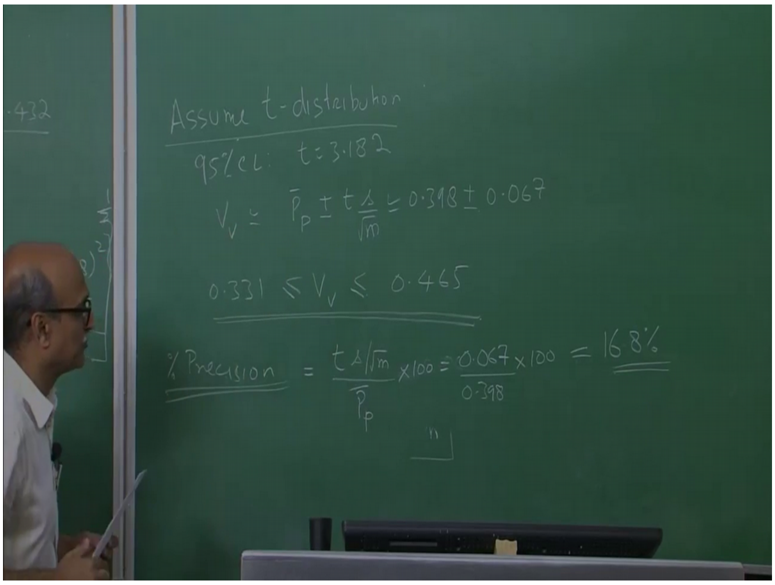
So, the mean value of the point fraction is simply add up all the 4 point fractions I have got 0.346 plus 0.432 plus 0.383 plus 0.432 divided by the number of measurements and that have 4. And this gives me 0.398 or simply 39.8 percent as the volume fraction or volume percent. Now this is a point estimate. And as I said we want to associate what is the kind of uncertainty that is associated with it. So, we calculate the standard deviation, s now the standard deviation s would be 1 upon m minus 1, well here m is equal to 4.

So, this is 4 minus 1 times now what I will have to do is add up the squares of the deviations from the mean value. So, it will be 0.346 minus 0.398 square, plus the second value 0.432 minus the mean which is 0.398 square plus third one 0.383 minus 0.398 square plus 0.432, well this is just the repeated result 398 square and whole thing squared to power half. So, you take the square root of the whole thing. Once you do this I get a standard deviation up to 3 decimal places at 0.042. Now suppose I want to find a 95 percent confidence limit. And let me assume normal that the sampling distribution is normal.

So, my volume fraction estimate would be. So, volume fraction range that I expect at the would be plus or minus z times s divided by square root m. And if I want 95 percent confidence interval c I I must take value of z is 1.96. So, it is easy now I have I know what is s I have already calculated which is 0.042, z is 1.96 m is 4. Hence this gives me instead of P bar now I can write 0.398, 0.398 plus or minus and calculate this. I will get 0.041, which translates to this is the range for my volume fraction or this is the error bar or the error interval, which becomes equal to 0.357.

So, 0.398 minus 0.041 will give me 0.357 to 0.398 plus 0.041 that gives me 0.439. So, what I am saying here that at the 95 percent confidence level, the range over which I expect the volume fraction of the graphite nodules for the particular sample would lie between 35 or 36 roughly 36 percent to about 44 percent. But we are assumed normal distribution; however, we have very few points just 4 points. So, it may be more prudent to assume a t distribution.

(Refer Slide Time: 31:28)


Assume t-distribution
95% CL: t = 3.182
$V_V \simeq \bar{P}_P \pm t\frac{s}{\sqrt{m}} \simeq 0.398 \pm 0.067$
$0.331 \leq V_V \leq 0.465$
% Precision $= \frac{ts/\sqrt{m}}{\bar{P}_P} \times 100 = \frac{0.067}{0.398} \times 100 = 16.8\%$

So, assuming a t distribution if I look at the t value for my m equal to 4. So, m is equal to 4 means the degree of freedom is m minus 1. So, 4 minus 1. So 3 So, 3 is the number of degrees of freedom, and hence the t value is 3.182.

So, for 95 percent confidence level t is 3.182. And from this I can now get a range for volume fraction as this and instead of basically writing z I write plus or minus t s upon square root of m. Now this gives me well the mean value of P bar P p bar is 0.398 plus or minus t is 3.182 s is 0.042 and m is 4. This gives me 0.067. So, what I will have for the uncertainty interval for the volume fraction s 0.398 plus or minus 0.067 that gives me 0.331, and the upper limit is 0.398 plus 0.067 and that gives me 0.465.

So, if I compare this particular interval against the interval where I had assume z, one would find that a t distribution would give you a larger interval, but if you have So, few data points it is best that we assume a t distribution to get the confidence interval. Another thing that we can estimate out of this is precision, what is the percentage precision we have? The percentage precision can be estimated let us say we are assuming a t distribution as simply t s divided by square root m this is my deviation around the mean divided by the mean value itself. So, this would give me 0.067 and then I needed in percentage. So, multiplied by 100. So, this would give me 0.067 divided by the mean P p bar which is 0.398 multiplied by 100 and this would give me a precision of 16.8 percent.

Now, this perhaps may be too large we want better precision than this, and hence what I suggest is that you look at the data not just a 4 points that we considered, consider for all the 20 points together. So now, how your m value will be 20, and if your m value is 20 if you go back to the t table for a value of 20 means degrees of freedom is 19, you will have a t value of 2.093. So, one is the t value would be different when I consider all the 20 data points here, the mean value would also come out to be different out of all the 20 values. Together and the standard deviation would also come out to be different as a result you would find that the you would get a much better or a smaller interval using the t distribution.

So, this would be given to you as an assignment problem also, but here are all the data is given in this particular lecture only for you to do this simple calculation. And just to I had given you how to do this all these calculations manually and I suggest you do these calculations manually; however, if I want to do the same estimates of these 4 values using this spreadsheet here. Then I will get let us say I want to get let me just copy these 4 values because these are the values that I had used. And take it here let me put it put these values here. So, this is small n out of that get point fraction which is simply 28 and remember there were 81 grid points. So, j 2 divided by 81 that gives me a point fraction from here and then just copy this formula and let me just put everything down to 3 decimal places.

So, I get a set of values which are the same what I had done on the board. Now I can get an average value as simply using the average function. So, this gives me mean then let me calculate standard deviation, which is again this spreadsheet has a function standard deviation and take out these values and one would get the standard deviation of 0.042 as what we had just seen. Now suppose I want to get a t distribution out here using the t statistic getting the confidence interval. Well, excel directly has a function I can go to formulas insert function go to statistical. So, this particular version of excel does not have for the t distribution.

So, let us just skip this I calculate now t s divided by square root of m. So, remember t was 3.182 multiplied by standard deviation which is k 8 divided by m is 4 square root of 4 is 2. So, I will just put 2 here and I get 0.067 is what I had calculated for the t. If I had done for the normal then it will be z s divided by square root m and this would be now instead of using t as 3.182 I am going to use z as 1.96 multiplied by standard deviation divided by again square root of 4 2 and that gives me 0.041. So, in excel also you can calculate these intervals quite easily.

So now you can do addition and subtraction for the mean to get the interval. So, this lecture we have done for some simple measurements, and how to estimate the uncertainty associated with the measurements. And with this, this lecture comes to an end.