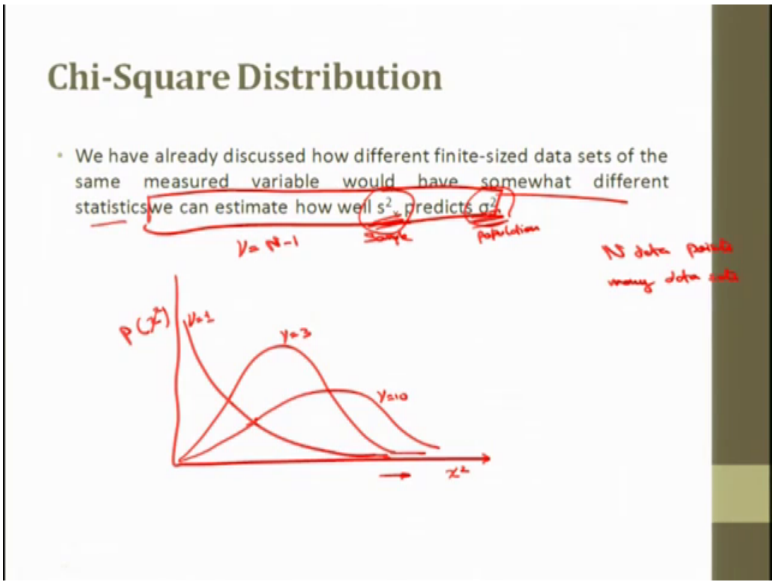
Engineering Metrology

Prof. J. Ramkumar
Dr. Amandeep Singh Oberoi
Department of Mechanical Engineering & Design Programme
Department of Industrial & Production Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture - 45
Chi square distribution, and Data outlier detection

Good morning. Welcome back to the course Engineering Metrology and I am taking statistics in metrology part in this course. So, this lecture we will continue the probability distributions, next I will move forward to Chi-square distribution.

(Refer Slide Time: 00:24)



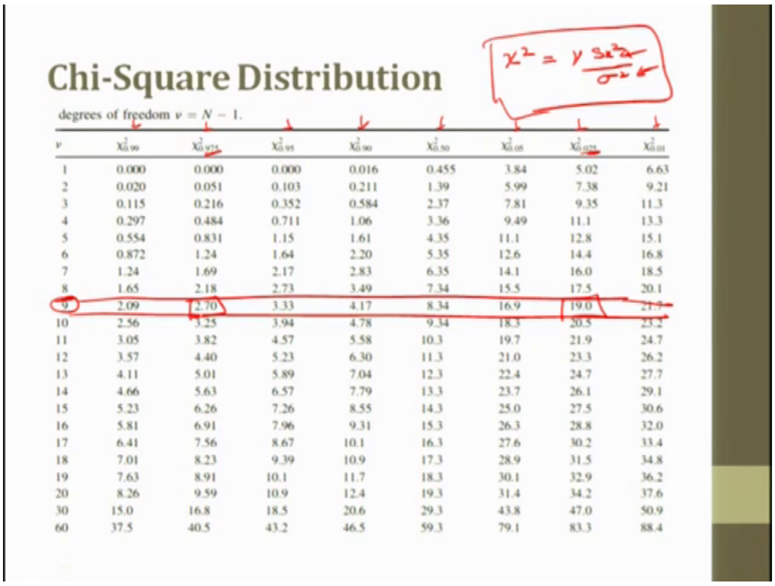
Chi-Square Distribution

- We have already discussed how different finite-sized data sets of the same measured variable would have somewhat different statistics we can estimate how well s^2 predicts σ^2

What is Chi-square distribution? We have already discussed how different finite sized data sets of the same measured value would be somewhat different statistic. We can estimate how well the standard deviation of the sample predicts the standard deviation of population, this is for sample; this is for population. So, when we do not know the value of mean, when some value of mean is not given and we are just working with standard deviations the Chi-square distribution is more useful. So, it because it can make the estimates based upon the standard deviation of samples only.

So, if we plot the sample standard deviations for many data sets each having N data point let me call it N data points, there are many data sets. We would generate the probability density function of p Chi-square probability density function of p Chi-square is again skewed like this probability of Chi-square. And it depends upon the degrees of freedom, degrees of freedom plays a great role here degrees of freedom is actually when we have N data points and data sets degrees of freedom is equal to V is equal to N minus 1. And when this N minus 1 value is one the distribution is somewhat like this, this is equal to this is for V is equal to 1. When this is equal to 3 that is for samples this is for V is equal to 3 and we have equal to 10 somewhat like this, this is above V equal to 10. So, this is probability for Chi-square and this is Chi-square. This is Chi-square in this direction.

(Refer Slide Time: 02:53)



Chi-Square Distribution

$\chi^2 = \nu \frac{s_x^2}{\sigma^2}$

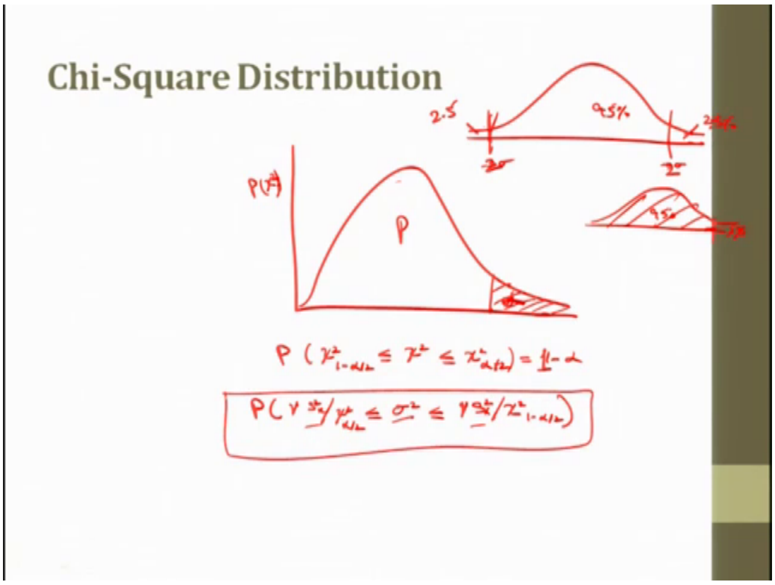
degrees of freedom $\nu = N - 1$.

ν	$\chi^2_{0.99}$	$\chi^2_{0.975}$	$\chi^2_{0.95}$	$\chi^2_{0.90}$	$\chi^2_{0.50}$	$\chi^2_{0.05}$	$\chi^2_{0.025}$	$\chi^2_{0.01}$
1	0.000	0.000	0.000	0.016	0.455	3.84	5.02	6.63
2	0.020	0.051	0.103	0.211	1.39	5.99	7.38	9.21
3	0.115	0.216	0.352	0.584	2.37	7.81	9.35	11.3
4	0.297	0.484	0.711	1.06	3.36	9.49	11.1	13.3
5	0.554	0.831	1.15	1.61	4.35	11.1	12.8	15.1
6	0.872	1.24	1.64	2.20	5.35	12.6	14.4	16.8
7	1.24	1.69	2.17	2.83	6.35	14.1	16.0	18.5
8	1.65	2.18	2.73	3.49	7.34	15.5	17.5	20.1
9	2.09	2.70	3.33	4.17	8.34	16.9	19.0	21.7
10	2.56	3.25	3.94	4.78	9.34	18.3	20.5	23.2
11	3.05	3.82	4.57	5.58	10.3	19.7	21.9	24.7
12	3.57	4.40	5.23	6.30	11.3	21.0	23.3	26.2
13	4.11	5.01	5.89	7.04	12.3	22.4	24.7	27.7
14	4.66	5.63	6.57	7.79	13.3	23.7	26.1	29.1
15	5.23	6.26	7.26	8.55	14.3	25.0	27.5	30.6
16	5.81	6.91	7.96	9.31	15.3	26.3	28.8	32.0
17	6.41	7.56	8.67	10.1	16.3	27.6	30.2	33.4
18	7.01	8.23	9.39	10.9	17.3	28.9	31.5	34.8
19	7.63	8.91	10.1	11.7	18.3	30.1	32.9	36.2
20	8.26	9.59	10.9	12.4	19.3	31.4	34.2	37.6
30	15.0	16.8	18.5	20.6	29.3	43.8	47.0	50.9
60	37.5	40.5	43.2	46.5	59.3	79.1	83.3	88.4

Now, what we are interested in? We are interested in finding the value of Chi-square and to just we need to see that is the data or the sample that we are talking about does that sample that the standard deviation predicts the population (Refer Time: 03:07) deviation or not. This is actually the key to use Chi-square distribution. Does the sample standard deviation predict the population standard deviation or not? Ok.

Now, this is the distribution of Chi-square in a tabular form we have degrees of freedom V is equal to N minus 1, and Chi-square can be calculated as a ratio of the sample and the population variances. This is equal to mu s x square over sigma square. Please be careful here that I am putting sigma for population note for sample here for sample I have put the value as x the sigma is the populations standard deviation and s x is the sample standard deviation.

(Refer Slide Time: 04:05)



Chi-Square Distribution

$$P\left(\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}\right) = 1-\alpha$$

$$\boxed{P\left(\nu s^2/\chi^2_{\alpha/2} \le \sigma^2 \le \nu s^2/\chi^2_{1-\alpha/2}\right)}$$

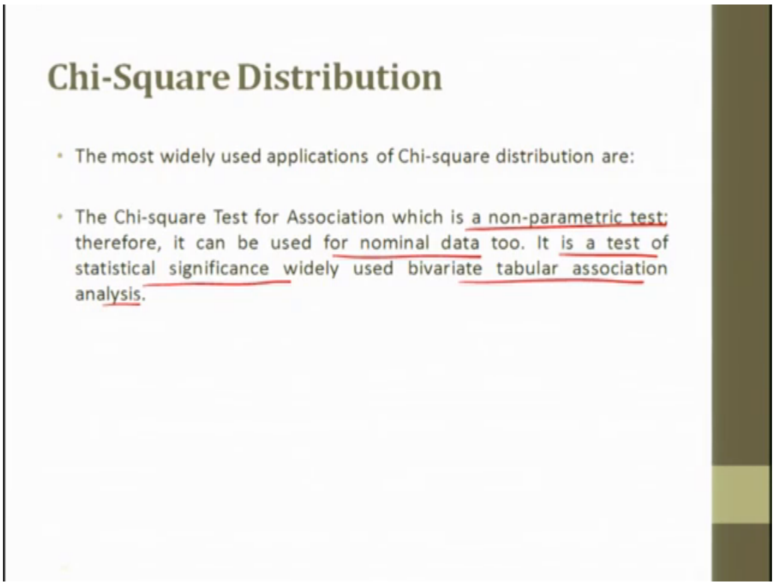
I am interested in finding the value of the probability rather is my sample significant or not. Now, I am talking about the word significant here. So, this I would like to talk in normal distribution as well I discussed this before also. When we have let me say 3 sigma minus 3 sigma and plus 3 sigma we had 99.7 percent area here, or let me call it 2 sigma, 2 sigma I have 95 percent area here to keep it a little simpler. So this, this 2.5 percent here and 2.5 percent here.

So that means, when I am selecting 2 sigma limits for my 2, see the credibility of my sample 2 sigma limits means I am 95 percent confident, if it is giving 3 sigma limits I have 99 percent confident. If I have 95 percent confident within 2 sigma limits that means, my significance level in case of two tailed test you have you can considering two tail it is 2.5 percent, in case of one tailed test we need not to go into the depth of these things in case of one tailed it is somewhat like this, if this is 95 percent this 5 percent is the significance level.

So, Chi-square test is used to find that whether the standard deviation of the sample predicts the value of standard deviation of population or not, that is also based upon the significance level. This value is known as is denoted as alpha here. So, this is the probability I have probability for Chi-square here. So, this is skewed data. So, I will try to solve this using a numerical using a problem here. So, it is a procedure interval for Chi-square I can just put it a probability for Chi-square 1 minus alpha by 2 less than equal to value of Chi-square with less than equal to Chi-square alpha by 2 which is equal to 1 minus alpha 1 minus alpha is that this probability this probability I am talking about this big area, ok, this big area.

So, if I put this relation this relation here and put the value of variance, so then I can put the Chi-square probability for the variance would be there then ratio of the degrees of freedom into variance for sample over Chi-square value Chi-square value for alpha by 2. And this is again on this side also I can have degrees of freedom variance sample divided by Chi-square for 1 minus alpha by 2, this is a general formulation. So, the value of alpha if it is let me say 5 percent alpha by 2 will be 2.5 percent. So, we can consider it for any limits.

(Refer Slide Time: 08:15)



Chi-Square Distribution

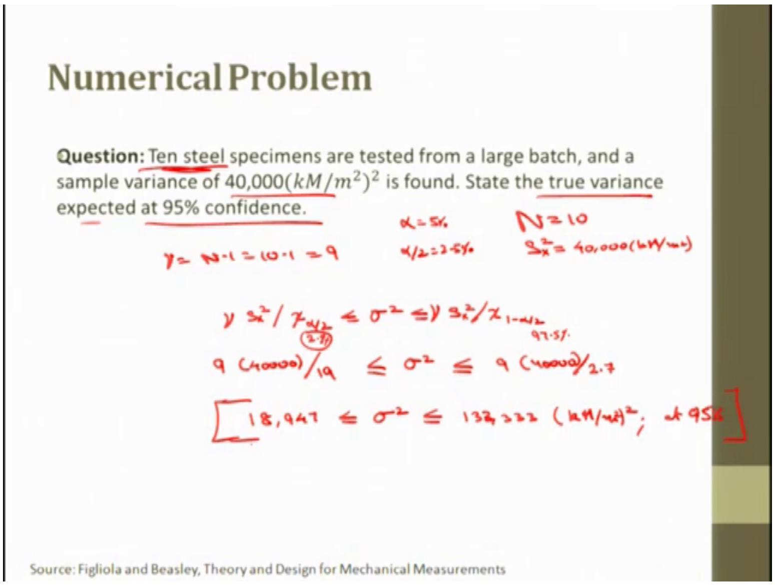
- The most widely used applications of Chi-square distribution are:
- The Chi-square Test for Association which is a non-parametric test; therefore, it can be used for nominal data too. It is a test of statistical significance widely used bivariate tabular association analysis.

So, before taking the example I like to tell you the significance of Chi-square distribution. Chi-squares distribution most widely used for the test of associations which is a nonparametric test; therefore, it can be used for nominal data two. It is a test of statistical signification significance widely used bivariate tabular association analysis.

So, typically when we have a hypothesis where we whether we need to see that whether or not, the two populations are different in some characteristic or aspect or the behavior is based upon two random samples this test procedure is known as Pearson Chi-square test. A Chi-square goodness fit of test is also used which is an observed distribution that confirms at to any particular distribution. Calculation of this goodness of fit test is by comparison of the observed data with a data expected based upon particular distribution. So, we will not move to that tangent here, but just we need to compare two samples, two samples based upon their standard deviation or we need to compare our sample to the population Chi-square test is used. This is the nutshell, this is the something in nutshell that we can there is a take away for you people.

So, Chi-square is using the standard deviation to compare two samples sometimes or to compare actually to compare two samples f test is there. This is for comparing the sample standard deviation to the population standard deviation.

(Refer Slide Time: 09:44)



Numerical Problem

Question: Ten steel specimens are tested from a large batch, and a sample variance of 40,000$(kM/m^2)^2$ is found. State the true variance expected at 95% confidence.

$\alpha = 5\%$ $\quad N = 10$

$\nu = N - 1 = 10 - 1 = 9$ $\quad \alpha/2 = 2.5\%$ $\quad S_x^2 = 40{,}000(kM/m^2)$

$$\nu S_x^2/\chi_{\underset{(2.5\%)}{\alpha/2}} \leq \sigma^2 \leq \nu S_x^2/\chi_{\underset{97.5\%}{1-\alpha/2}}$$

$$9\,(40000)/19 \leq \sigma^2 \leq 9\,(40000)/2.7$$

$$\left[\,18{,}947 \leq \sigma^2 \leq 133{,}333\ (kM/m^2)^2;\ \text{at } 95\%\,\right]$$

Source: Figliola and Beasley, Theory and Design for Mechanical Measurements

So, I have at this problem here the 10 steel specimens are tested from a large batch, 10 steel table is made from large batch and a sample variance is this. State the true variance expected at 95 percent confidence?

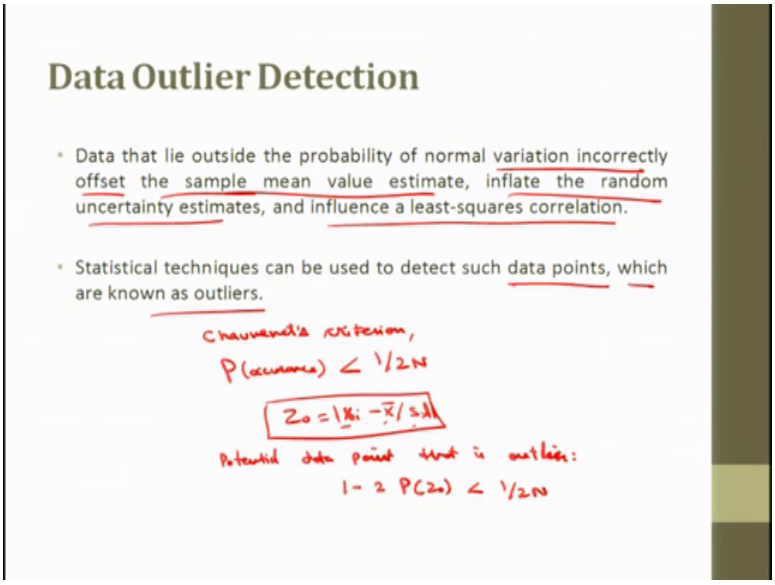
If this kind of problem is there we know the sample variance only and we know how many specimens are selected we just know the value of N is equal to 10 and we know the value of s x that is equal to s x square actually, s x square actually 4000 kilometer per meter square. So, for this I have degrees of freedom is equal to N minus 1 is equal to 10 minus 1 is equal to 9, and he has called it 95 percent confidence interval, for 95 percent confidence interval I will take the value of alpha. So, here the value of alpha is 5 percent and alpha by 2 is 2.5 percent.

So, if I use this relation, this relation to solve this kind of problem I can just put here mu into the variance for the sample by the Chi-square value for Chi-square value for alpha by 2 is less than equal to sigma square will less than equal to degrees of freedom into variance for the sample divided by Chi-square values for 1 minus alpha by 2, ok. I can just put the values directly here it is 9 into 40,000 divided by Chi-square value I will just let you see how to find the Chi-square value here less than equal to this, this is less than equal to 9 into 40,000 divided by Chi-square for; this is Chi-square for 2.5 percent actually this is Chi-square for 91 minus alpha by 2 that is 100 minus alpha by 2 actually 0.25, this is Chi-square for 97.5 percent, ok.

Now, let us see the Chi-square table. The value first is we need to select this degrees of freedom is 9. So, this is the row that is selected. So, we have Chi-square distributions for different levels of significance here. So, for 0.025 the value is 19, and for 97.5 the value is 2.7. So, these two values are there this value is 19 for 0.025 it is 2.5 percent for 97.5 percent value is 2.7. So, let me try to put these values here.

For this value is 19, for this the value is 2.7. Here said it state the true variance expected. So, the precision interval for the variance is if I calculate this, this is 18,947 is less than our variance the precision variance for the population and if this is less than equal to 133,333 I can put the units for the variance as well kilo m plus by m square whole square because these variance have taken square at 95 percent, ok. But these precision interval about the variance due to random chance then as N becomes larger the precision interval narrows as s approaches our like sample approaches the population. So, this is the way to find the precision interval for our variance if it is required.

(Refer Slide Time: 14:36)



Data Outlier Detection

- Data that lie outside the probability of normal variation incorrectly offset the sample mean value estimate, inflate the random uncertainty estimates, and influence a least-squares correlation.
- Statistical techniques can be used to detect such data points, which are known as outliers.

Chauvenet's criterion,

$P(\text{occurrence}) < 1/2N$

$\boxed{Z_0 = |x_i - \bar{x}/ s_x|}$

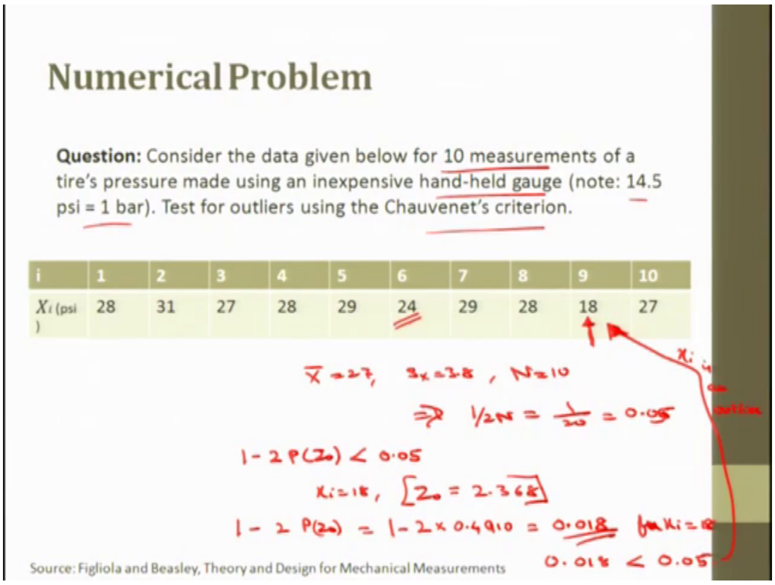
Potential data point that is outlier:

$1 - 2\,P(Z_0) < 1/2N$

Next and the last part in this presentation is data outlier detection. Sometimes we have data outliers which are due to some assignable causes or which are due to some non-random causes. So, we need to find out but where is the data outlier; there is a rule to find that. So, data that lie outside the probability of normal variation in correctly offset the sample. Sample mean value estimate inflate the random uncertainty estimates and influence a least square correlation. So, test statistical techniques can be used to detect such data points which are known as outliers. So, one approach to detect the outlier is Chauvenet's criterion, ok.

This identifies the outliers having probability less than 1 by 2 N. The probability of occurrence is less than this. So, to apply this criterion we need to just need to find the value of Z naught, which is equal to the modulus of x i minus mean as we did it before. So, Z naught is equals x i minus x bar by standard deviation of x and the probability values for the data point through the potential data point that is outlier; I can put here potential data point that is outlier that is calculate calculated as 1 minus twice of the probability for the Z naught if it is less than 1 by 2 N.

(Refer Slide Time: 16:39)



Numerical Problem

Question: Consider the data given below for 10 measurements of a tire's pressure made using an inexpensive hand-held gauge (note: 14.5 psi = 1 bar). Test for outliers using the Chauvenet's criterion.

i	1	2	3	4	5	6	7	8	9	10
X_i (psi)	28	31	27	28	29	24	29	28	18	27

$\bar{x} = 27$, $s_x = 3.8$, $N = 10$

$\Rightarrow 1/2N = \frac{1}{20} = 0.05$

$1 - 2P(z_0) < 0.05$

$x_i = 18$, $\boxed{z_0 = 2.368}$

$1 - 2P(z_0) = 1 - 2 \times 0.4910 = \underline{0.018}$ for $x_i = 18$

$0.018 < 0.05$

Source: Figliola and Beasley, Theory and Design for Mechanical Measurements

So, let me try to see this problem here. So, we have 10 measurements of a tire pressure made using an inexpensive hand held gauge test for the outliers using Chauvenet's criterion 10 measurements, and we have this observation table here. Now, what we need to do to calculate these?

We need to first find x bar s x, x bar, x bar is nothing but sum of these values and average of these values. So, x bar that is calculated to be x bar is 27 here and s x is 3.8. So, let me try to apply that criteria and N is equal to 10, ok. Now, N is equal to 10 that implies 1, 1 by 2 N is equal to 1 by twenty is equal to 0.05 this means the outlier would be if 1 minus 2 into probability of z naught is less than 0.05, ok.

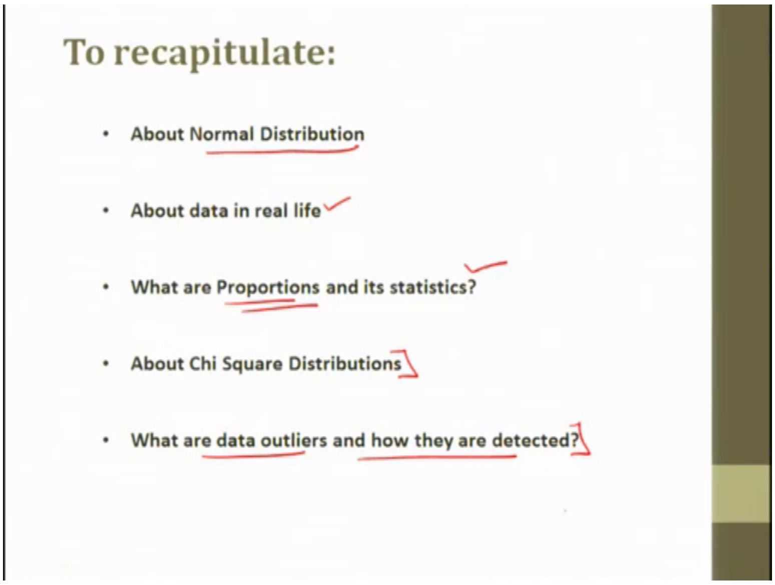
Now, we can calculate the value of Z naught also using the relation that was given we can find for the 10 data points; x i is actually our potential outlier I can check it for this value, ok. So, this is x i that is selected fix, x i is selected here x bar I know. So, 18 is my potential x i and Z naught I can find it using this value for that means, for a x i is equal to 18 if I see my table I will find and using the relation given in the previous slide I can find the value of Z naught that is equal to 2.36 8, ok.

So, 1 minus 2 into probability of Z naught which is equal to 1 minus 2 into from the table if I see probability of Z naught would come out to be 0.4910, Z naught is calculated using this relation, ok. This is the modulus here z naught will always positive here. So, this probability is equal to 0.018, ok. Now, this is the calculated probability for my x i

18, this is for x i is equal to 18 and this is my 1 by 2 N. So, I can see here that 0.018 is less than 0.05 and now this implies that this x i, x i is an outlier, ok. So, x i is the outlier, 18 is the outlier here. Also we can check another outlier we like we if we interested we can even check whether 24 is outlier here.

So, in this lecture it was might be quite heavy for you because we had new concepts and we have many mathematical relations, but do not worry we will provide you with the notes to read more on this like Chi-square distribution, what are the various applications and we will put a quiz. And please do ask the logical questions as well in the forum. I will be very happy to respond to them. And we will meet in the next lecture.

(Refer Slide Time: 20:43)



To recapitulate:

- About Normal Distribution
- About data in real life
- What are Proportions and its statistics?
- About Chi Square Distributions
- What are data outliers and how they are detected?

Just to recapitulate I discussed the normal distribution, then about the application of normal distribution in real life, I did some numerical. Then I discussed about the proportions and binomial distribution, then we had some information on Chi-square distribution to compare the variances of the sample and the population, then we saw a rule to detect the outliers.

So, we will meet in the next lecture. We will discuss the statistical parts in metrology further.

Thank you.