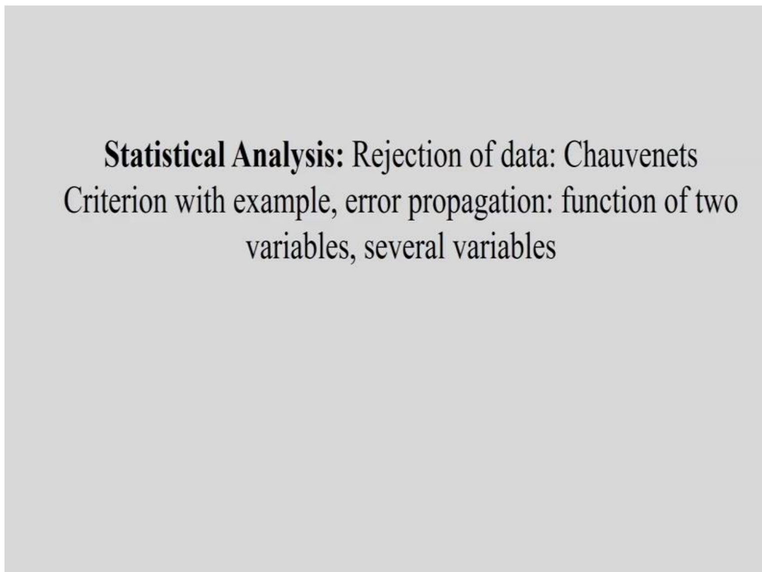
Experimental Methods in Fluid Mechanics
Professor Dr. Pranab Kumar Mondal
Department of Mechanical Engineering
Indian Institute of Technology, Guwahati
Lecture 41
Rejection of Data- Chauvenets Criterion with Example

(Refer Slide Time: 00:30)



Good morning, we will discuss today the rejection of data, that means in this is another important module of this course that the data which are collected from the experiments, we need to know whether the data which are collect collected are correct or not and rejection of data is another important module of this course.

So, before you go to discuss about the rejection of data will try to discuss another important aspect that we could not discuss in the last class that is the, measurement error. So, if we try to recall in the last class, we have discussed about the, that sample variability and the measurement error which are the two important reasons for the statistical I mean, these two are the important reasons for which we need to go for the statistical analysis.

If we take that example again that if we if a particular production unit is developing a particular new plastic and if we take a 100 samples from the production line, and if you try to measure the ultimate strength, then what we will get we will get 100 different values for the 100 different samples.

Now, we have discussed that there are two different reasons for which we will get a 100 different values for the 100 different samples. One is the sample variability another is the measurement error and we have discussed that the sample variability regardless of the sample variability we need to describe that I mean, we would be compelled to describe it and the description which is done using the statistical analysis.

In continuation of that, today we will discuss another important point which is required to estimate and in fact, which is required to know whether the measured value or measure data is correct or not and for that if we consider another example, say one manufacturing unit is producing gas tank and if you are responsible for checking the safety, I mean when the manufacturing unit is producing gas tank before the product goes into the market.

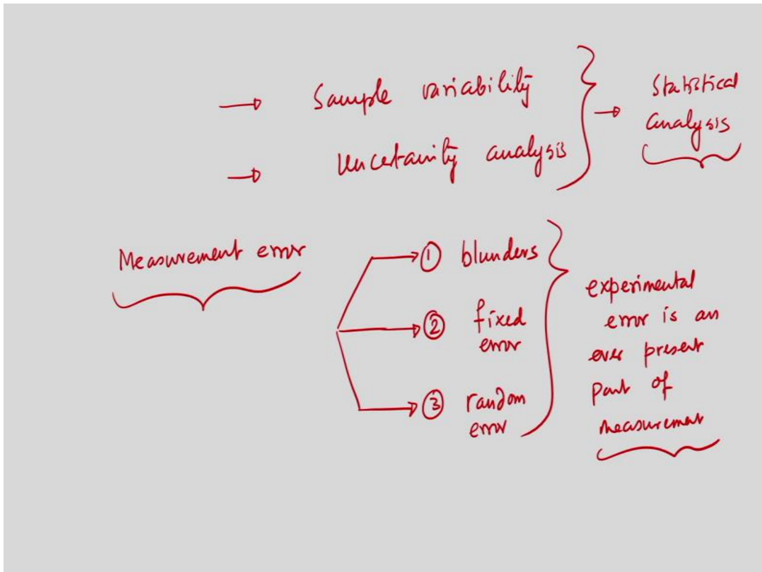
Our objective should be to certify whether the product will be safe or not. That means, before the product is taken for the practical application, we need to ensure we need to certify that the product will be safe or not. Now, since the product is gas tank, so the internal pressure of the gas is an important aspect to know.

So, when the tank is producing now whether the gas tank will be able to stand particular pressure or if we increase so, I mean when a particular product is developed say gas tank is manufactured now, the gas tank is manufactured keeping in mind that there will be a design value of the internal pressure that the gas tank will be able to stand.

Now what will happen if the internal pressure of the gas increases or becomes 10 percent higher of the design value, then what important question will be will the tank be able to stand the gas gas pressure. If yes, then fine if it is not then why, so that means whether it is 10 percent higher of the design value or sometimes if it becomes 15 percent higher of the design value, so that is what we need to specify and that is another important aspect of this module is that I am talking about that the uncertainty analysis.

That means, we really do not know whether the gas tank that is manufactured in this manufacturing unit and if we take samples of the production line and if we randomly measure the internal pressure that whether the gas tank will be able to withstand the pressure or not and we need to certify we need to ensure what will be the uncertainty uncertainty range, that means if the design value is say x, then x plus minus 5 percent or 10 percent will be able to withstand the gas pressure or not. If the design value is x and if the gas pressure is less than that value, no problem and but if it becomes higher than the design value then what will be the range over which the tank able to withstand the pressure and this is known as uncertainty analysis.

(Refer Slide Time: 05:41)



So we have seen that sample variability and second is uncertainty analysis. These two if we need to ensure, we need to go for the statistical analysis. So that is why statistical analysis is important to check the safety of the product and to certify the range over which the product will be able to work without any failure.

So now, these two are important second is the measurement error that is what we have discussed measurement error. Now again, if we consider that example that production unit is developing new plastic and if we take samples, I can say that if we take samples of the production line say production line, so if we take 100 samples of the production line, and if we measure the rupture strength or the ultimate strength, then what we will see, we will get 100 different values. One reason is the sample variability that is what we have discussed.

Another reason will be the measurement error, that means whether when we are measuring the ultimate strength using any I think using in a in particular instrument or measuring system, then whether the measured value we are recording we are tabulating is correct or not that means weather during the measurement we had measurement error or not.

So, we need to know, what is measurement error? That means when we are measuring using any device using an instrument on measurement system, we really do not know whether the error is associated the measurement or not. So maybe we are getting 100 different values for 100 different samples. One reason is the sample variability other reason could be measurement error and the measurement error is also an important part that will should look at.

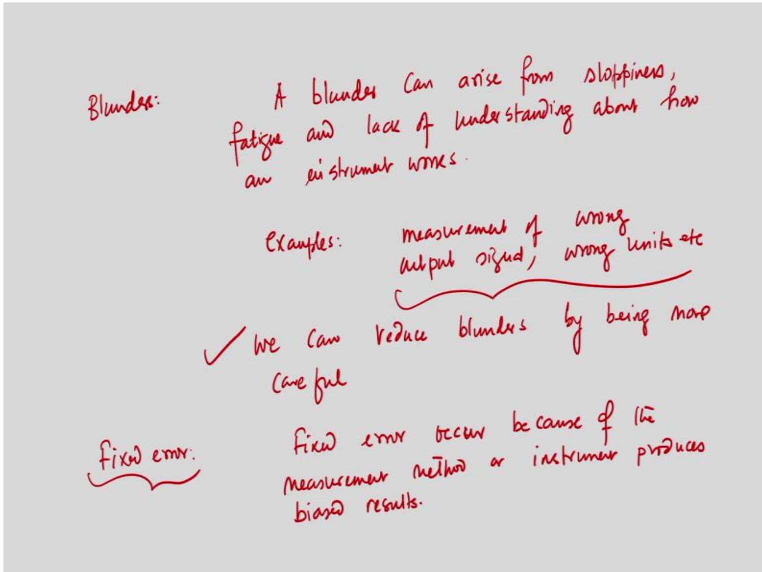
So the measurement error, there are three different errors, one is known as you know that is called this blunders, number 2 is fixed error. Number 3 is random error, fixed error another is the random error. So these three are the different errors we need to study. What is blunders? What is fixed error? And what is unknown error? So this is fixed error and this is random error.

So we really do not know the whether we are considering these three important points while measuring the ultimate strength of the plastic and may be because of this because of the fixed error or the random error, we are not getting the correct value. So what I like to say, the errors experimental error is an ever present part of the measurement and they will arise no matter how carefully we are taking the measurement.

So we cannot eliminate completely the measurement error, but our objective would be to minimize the error. So in line with the sample variability, measurement error is one of the important reasons for which we will get a 100 different values if we measured ultimate strength of a 100 different samples.

So this is important aspect, so what I was telling that the measurement error will arise no matter how carefully we are taking the measurement and that is why our objective should be we objective should be to minimize to reduce the measurement error, but our objective should not be to completely eliminate the error.

(Refer Slide Time: 10:39)



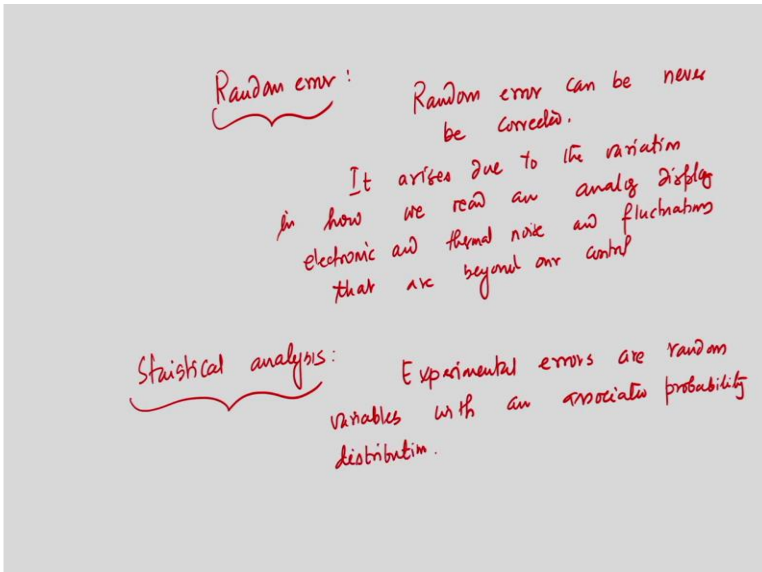
So, we will discuss now one by one, what is blunder? A blunder can arise from sloppiness, fatigue and lack of understanding about how an instrument works. So, example will be measurement of wrong output signal wrong units etc. That means, the blunder arise from the sloppiness, we are not careful in recording the value from the fatigue and also the lack of understanding about how are instrument works.

So what we can do, we can reduce blunders by being more careful. So this is important point, that means if we carefully, seriously measured the experimental data, then we can reduce the blunder. So this is blunder another is the fixed error. So this is blunders now fix error which I is important so I can say this fixed error occur because of the measurement method or instrument produces biased result.

If you are measuring distance using a metal tape, and if we measure the same distance using the same metal tape during winter and summer. So we are taking our first measurement during winter using same metal tape and the same distance were again measuring in summer, what will happen? Since you are using metal tape, so we should be careful that the there will be thermal expansion of the metal tape.

So if we do not take that effect into account will come up with the wrong results. So this is the fixed error. That means if we use metal tape in measuring distance, although the distance will remain fixed but if we take the measurements in summer, I mean we will get different value and if we take measurement using the same tape is in winter, we will get another value. So this is which is known as fixed error. So we will not be able to get the correct in our result unless we apply a temperature correction. So if we use this metal tape in measuring the distance, we need to consider the temperature correction only then I will get the correct value.

(Refer Slide Time: 15:07)



And next is the random that is another important that is the random random error. What is this? So, random error is another important part that is what I always thinking that as I said that we can minimize, we can reduce the error, we cannot eliminate completely. So random error is the error which can never be corrected. So random error can never be corrected, this error arises due to the variation of variation rather variation in how we read an analog display electronic and thermal noise and fluctuation that are beyond our control.

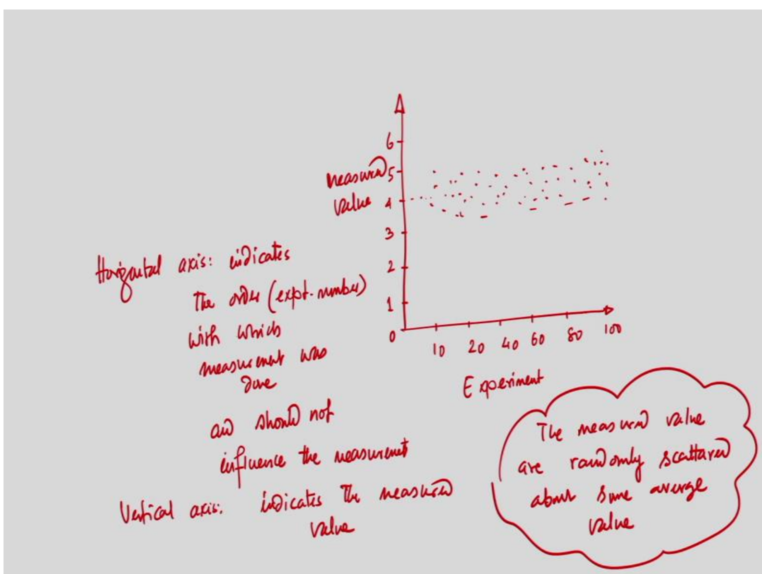
So which are beyond our control we cannot correct it, we cannot be we cannot correct this so these errors cannot be corrected which are beyond our control, so this is the random error. So what we have understood is that the sample variability and measurement error. These are associated with the measurement method, measurement technique, and we should be careful if we when you are measuring the when we measuring any parameter variable, then these are associated with the measurement process.

And considering this, if we consider 50 different samples 100 different samples will get 50 or 100 different values. So knowing that and knowing the fact that they are, are this two important reasons associated with the measurement process. That is the sample variability and the measurement error, we will not get correct value that means we really do not know I mean if you are taking 100 samples of the production line and if we measure maybe any variable whether it is ultimate strength, whether it is rupture strength, then we will get different values.

So next what we need to do is, we need to go for the statistical analysis. That means, if we get 100 different values than which is the correct one, should we consider the average of that and if we consider average then, I mean all the samples I mean all the measurements may be because of this error fixed error, random error and blunder all the measurements may not be or whether should not be the correct measurement.

So that means whether we need to eliminate a particular data from the measurement, I mean if we record 100 data points out of a 100 data points, we need to reject a few of a few data points and if after rejecting a few data points if you take average whether the average will be I mean we can certify the average value with the product and for that we need to go for the statistical analysis. So now statistical analysis. Why we need to go for that I am writing experimental errors are random variables with an associated probability distribution.

(Refer Slide Time: 20:13)



So if we go to the next slide and if we try to draw one curve, say we are drawing, this is experiments. So 10, 20, 40, 60, 80 and 100. So this is 10, 20, 40, 60, 80, 100. So this is the experiments. So we are doing a 100 experiments, we are doing 10 experiments, we are doing 20 experiments and this is measured value, So this is measured value. So that means the horizontal axis indicates I am writing this axis indicates the order that is experiments number experiments number with which measurement was done.

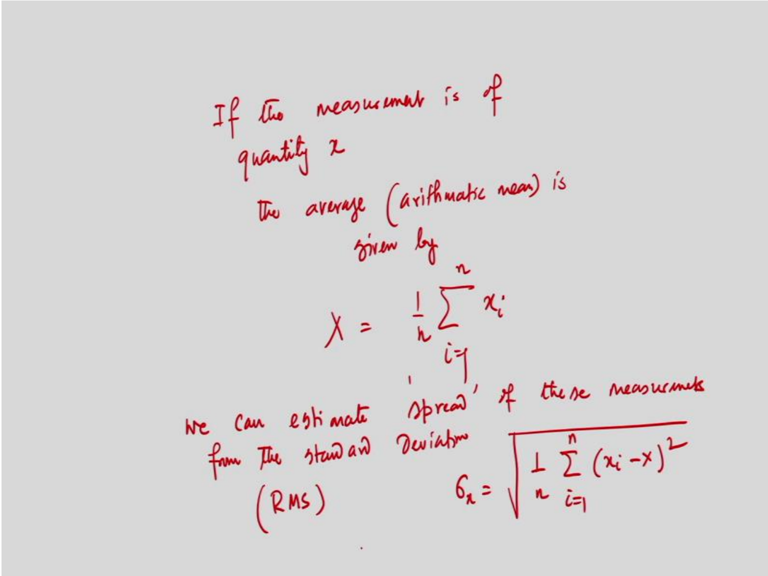
And should not influence the measurement and the vertical axis is and the vertical indicates the measured value. So along the horizontal axis, we are considering the order that is number of experiments we have conducted to measure apart measure to measure the value and should not influence the measurement.

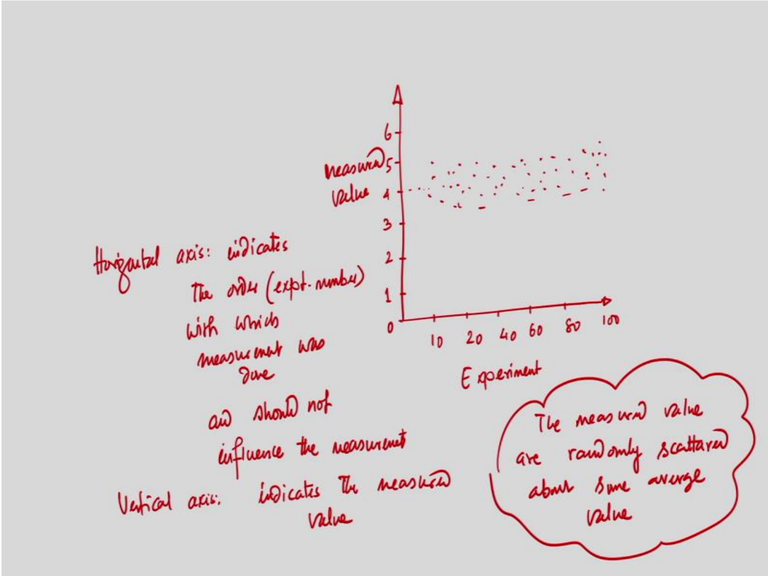
So that means if we this quite obvious if we are interested or if we are measuring a particular value, if we perform 10 experiments will get 1 value, if we perform 20 experiments we are supposed to get the same value because the the fixed data, so that is what I was telling if we measured distance using the metal tape, if we measure the same distance over (())(23:01) anytime we measured the distance were supposed to get the same value.

Now the we will not get the same value and that is what I was talking about. Now the measured values are randomly scattered. So if we now plot say this is 0, 1, 2, 3, 4 this is 5 this is 6 say this is we are getting scattered value.

So if we perform a 100 different experience will get scattered value. So what we can see from this figure that we are measured value the measured value or randomly scattered about what around the average value, that means if we look at the figure, what we can see if we perform 10, 20, 40, 60, 80, 100 experiments, we are supposed to get the same value, but we are not getting instead we are getting the measured values are scattered and randomly scattered about some average value, so this is important.

(Refer Slide Time: 24:47)





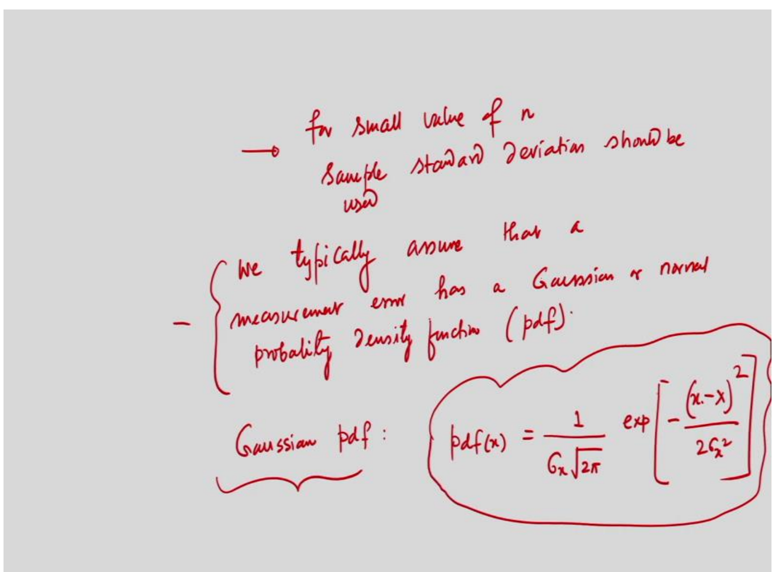
Now, if we go to the, rather if we can say if we go to the next slide, say the measured I am writing, if the measurement is of quantity x the average, that is arithmetic mean is given by x will be equal to 1 upon n, i equal to 1 to n xi. So if the measurement is a quantity x, so any quantity x we are measuring and if we vary n from 1 to 100 and we will take we will get the measured values are randomly scattered about some average value.

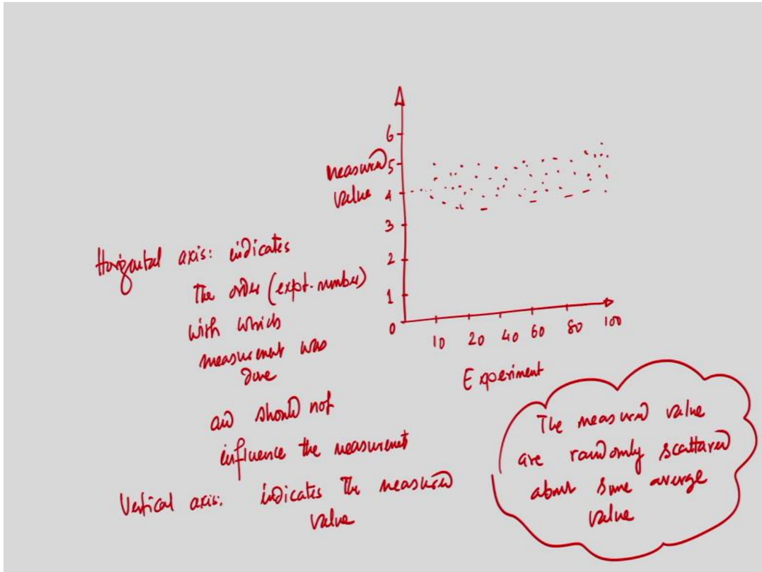
So we need to consider the average value and this is the average is the arithmetic average, that is what we have considered. And if we see the figure again, so if we look at that 15 to 20 percent measurement fall between 5, 4.0 and 4 point, so if it is correct, then only a few data if we look at this figure what we can see there are only 15 to 20 percent data, rather the data points I can say 15 to 20 percent measurements fall between 4 to 4 point 4 point 0 to 4 point 2.

So next our objectives should be that if we get this average value, then we can estimate the spread of the measurement from the standard deviation. So we need to go for the statistical analysis. We need to know what will be the error, that means if we we have seen that only 15 to 20 percent measurement fall between 4 and 4.0 to 4.2, remaining data points are very much scattered. Now we need to estimate the error of the measurement, so what we can do, we can estimate spread of these measurements from this standard deviation.

And that is sigma x will be 1 upon n summation of i equal to 1 to n, xi minus x square. So this is the spread of that is first standard deviation or sometimes it is known as root mean square. So this is called as root mean square value. So, this is what we can calculate we have seen that only 15 to 20 percent data roughly the, I mean 50 percent measurements fall between 4.0 to 4.2 remaining other measurement are discrete. So we need to know the spread of this measurement from the standard deviation.

(Refer Slide Time: 28:22)





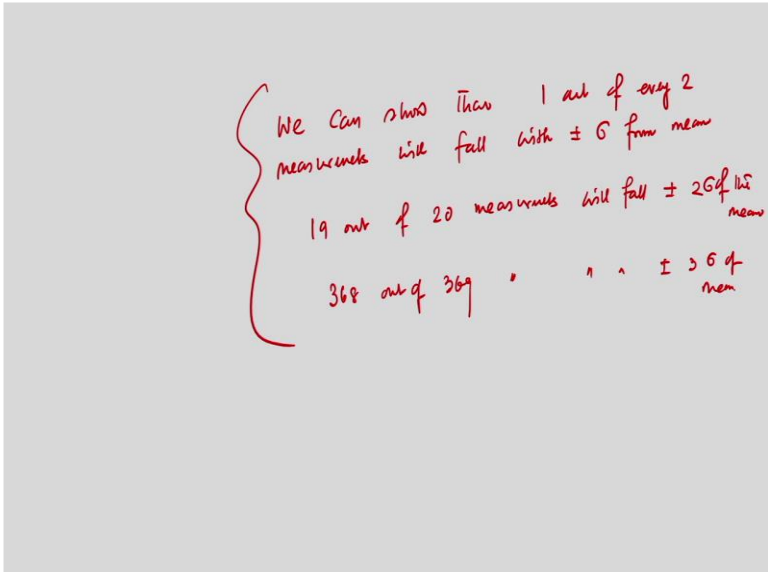
And that we can say for small value of n, the sample standard deviation should be used, sample I am writing sample standard deviation should be used. So what we can see? If we now go back to the previous slide what we can see. If we do 20 measurement, 40 measurements, 60 measurements or 80 measurement we can see that there is there is measurement there is a measurement error.

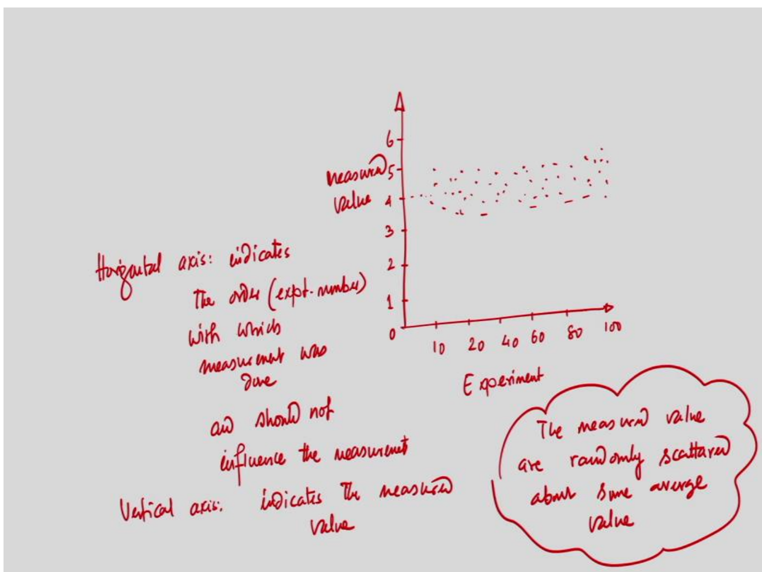
So that means, we cannot, redo we cannot eliminate so our objective should be to reduce the measurement error for that we need to go for the statistical analysis. What we will assume for this statistical analysis, we typically assume that measurement error has a Gaussian or normal probability density function probability density function (pdf).

So we typically assume that a measurement error has a Gaussian or normal probability density function normal pdf, but this is not in our case, which is you know, this is not always the case what rather we can that means if we assume that the measurement error will have a shape which is normal probability density function or Gaussian distribution.

This is not always the case rather our assumption behind this our assumption behind this shape should be verified. So, if we consider the Gaussian pdf, so what I am telling that this assumption is not always the case, what you need to consider the assumption of rather any assumption about the shape of the pdf should be verified. So if we consider the Gaussian pdf, then we write pdf function x that is 1 upon sigma x root 2 pi exponential minus x minus x any data square divided by 2 sigma x square. So that is what we can assume but again, I am telling that this is not the always case and the shape should be verified. Now, so what you can see that the even if we consider this the Gaussian distribution,

(Refer Slide Time: 32:25)





So what we can see that we can see that we can show that 1 out of every 2 measurements will fall within plus minus sigma from mean, 19 out of 20 measurements will fall plus minus 2 sigma of the mean and 360 out of 369 measurement will fall plus minus 3 sigma of mean. So this is what we can consider.

So now next is if we go to the this plot, so what we can see that the data points are randomly scattered and what I was telling that if we take 20 measurements or 40 measurements I mean of a the quantity x. The discrete data points because of the measurement error. Now, we need to a certain the range of uncertainty and if all the data points, I mean we can see that the data points are randomly scattered only 15 percent measurements fall between 4.0 to 4.2. So there will be a range of uncertainty we cannot consider a data point which is beyond that uncertain uncertainty range. So, that means the data points which we have recorded from the measurements measurement process, all the data points are not correct and we need to exclude a few data points, that is very important. That means we need to reject data from the tabulated data set.

So all the data points measured or collected from the measurements, measurement method are not correct rather we need to reject a few data points from the measured value and when we are going to reject any data, we should have we cannot arbitrarily reject any data. What we can see from the from the graph is that that only a few data points fall between the specified range remaining data points are not in the specified range.

And while we are trying to reject data which are collected or which are recorded from the measurement method, we need to go for certain procedure that means without I mean when you are planning to reject data point we should have our basis, so without having any basis we cannot reject data.

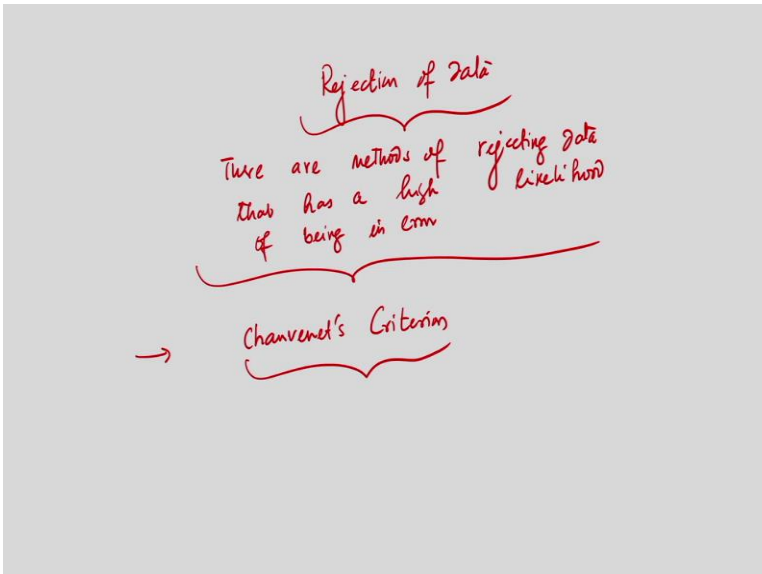
So that means, there is no absolute means of rejecting any data unless we specifically know it is wrong. So, we have gathered or we have collected we have collected 20 data points or 100 data points out of a 100 data points, maybe 50 or 60 or I mean 95 data points are correct and 5 data points are not correct.

If we know them the that those data points are wrong then only you can reject and we otherwise you cannot reject. While even if you know that 5 data points are wrong, then we can reject but but again while we are trying to reject data, we should have basis of reduction and there are methods of rejecting data and these data points, I mean which are going to reject that data has high likelihood of being in error.

So that means, if we know that the data point is going to be an error a particular data point which you have recorded. That data point is not correct one, that is because of the error then we can reject but again, we will have a particular method and following that method we can reject data.

So, when we are trying to reject data, we should have common basis and then only we can reject data.

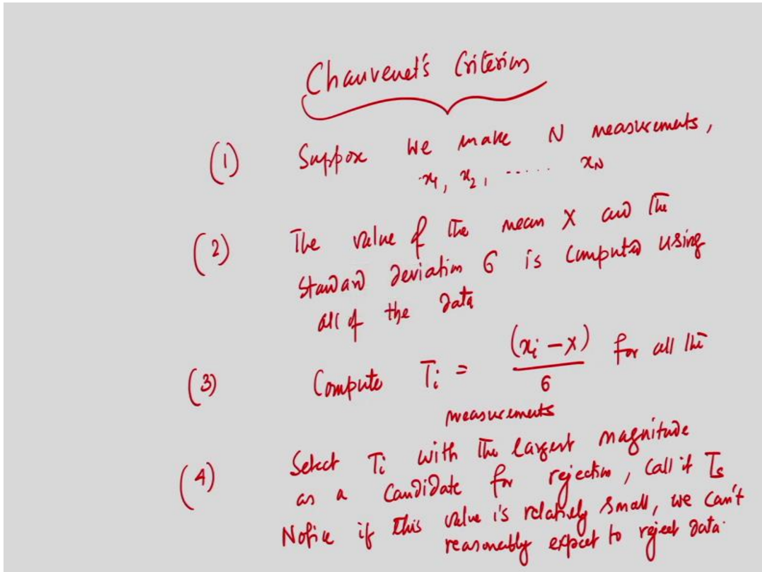
(Refer Slide Time: 37:54)

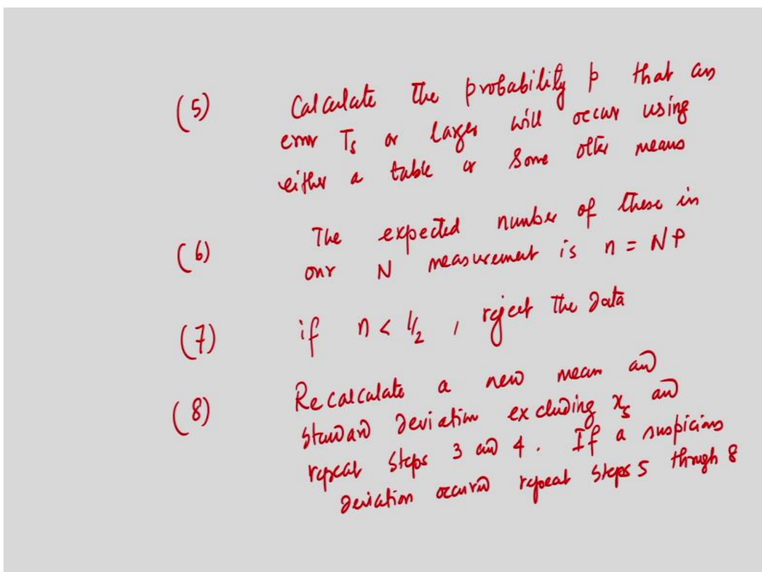


So rejection of data, what I said that there is no absolute means of rejecting data unless we specifically know that is wrong. If we know that is wrong then you can reject, but there are method. If we collect 100 data points, it is very difficult to say which data point is wrong. But even if we know that is wrong, then we can eliminate or we can rejected it.

But if we have collected 1000 data points, so out of 1000 data points a few will be definitely wrong, but we need to reject all the data points and to reject those data points there are a few methods. So there are methods of rejecting data that has a high likelihood of being in error and one method is known as Chauvenets Criterion Chauven it is Chauvenets Criterion. So Chauvenets Criterion is one of the methods following which we can reject data.

(Refer Slide Time: 39:47)





So you know, what are the points? I will be writing now the Chauvenets Criterion. So what is the Chauvenets Criterion? So Criterion is that number 1, so Chauvenet Chauvenets Criterion suppose I am writing the points and these points should be followed to reject a particular data and if we know that that particular data is wrong, suppose we have or we make 10 measurements like you know x1, x2 up to xn.

So we are really doing experiment and from the experiment we are taking N measurements like x1, x2, xn. Then the value of the mean X and the standard deviation sigma is computed using all of the data. So there this is number 2, so we are discussing today Chauvenets Criterion and the other steps we need to follow following this Criterion. Number 2 is this that will calculate mean and the standard deviation using all all of the data.

3, then we need to compute Ti that is xi minus x divided by sigma for all the measurements that is error we are calculating xi minus x divided by sigma. Now select Ti the largest magnitude as a candidate for rejection and call it Ts. Here, an important point to note is that notice if this value is relatively small, we cannot reasonably expect to reject data. If this value is relatively small we cannot reasonably expect to reject the data,

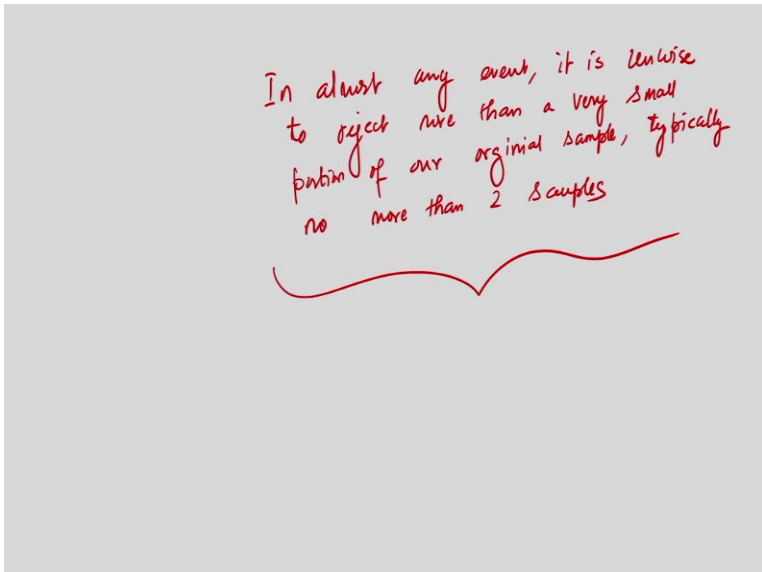
Then, what is the next step? Next step is 5 that calculate the probability P that and error Ts or larger will occur using either the table or some other means occur using either a table or some other means. Number 6 is, that means we need to calculate probability that an error Ts that is what we have calculated or larger will occur using either a table or some other means. 6 is the expected the expected number of these in our N measurements is n is equal to N into P.

7, if n is less than half, reject the data and then from the remaining data recalculate a new mean and standard deviation excluding xs and repeat steps 3 and 4. If a suspicious division occurred repeat steps 5 through 8. So if I go to the previous slide upper calculating Ti select Ti is a largest as a (())(46:25) rejection that fine, but if these Ti and that called Ts, but if this value is very relatively small then we cannot reasonably expect to reject data, then what we need to do we need to go steps 5, 6 and 7.

That means, we need to calculate probability that an error Ts larger will occur using either a table or some other means some other means is that the expected number of this order will be N into P and if n is less than half, then we can reject the data. And after rejecting a particular data we need to recalculate the new mean and standard deviation excluding that reject data that is excess and we can repeat steps 3 to 5, if we find at that you know that is Ts is larger than you can reject but if after following steps 3 and 4 again, if a suspicious deviation occurred we need to go for I mean repeat we need to repeat steps 5 through 8.

So this is basically the Chauvenets Criterion using Chauvenets Criterion of rejecting a data point and we have seen that all the data points we are collecting from the from any particular measurement, you know method may not be the correct one. So we need to go for the rejection and why when we are going for the rejection, we cannot arbitrarily rejecting data unless we know it specifically know it is wrong, but there are methods and we have discussed one of the methods is Chauvenets Criterion.

(Refer Slide Time: 48:02)



Finally, what I like to say, we will discuss this again take an example but in almost and any event, it is unwise to reject more than a very small portion of our original sample typically no more than two samples. So, this is very important point we should should keep in mind that in almost any event it is unwise to reject more than a very you know, various small portion of our original sample. Typically no more than two sample. So that means we can reject maximum 2 data points, we cannot reject more than that in most of the applications typically,

So what we will discuss this again taking an example that if you would like to reject data following the Chauvenets Criterion, then we will discuss will take a few data points and then if we know a priori that a particular data point is very large and then if we know that this is specifically wrong and we will check whether the Chauvenets Criterion is getting satisfied to reject the data not and we will do that in exercise in the next class. So, with this I stop my discussion today and we will continue in the next class. Thank you.