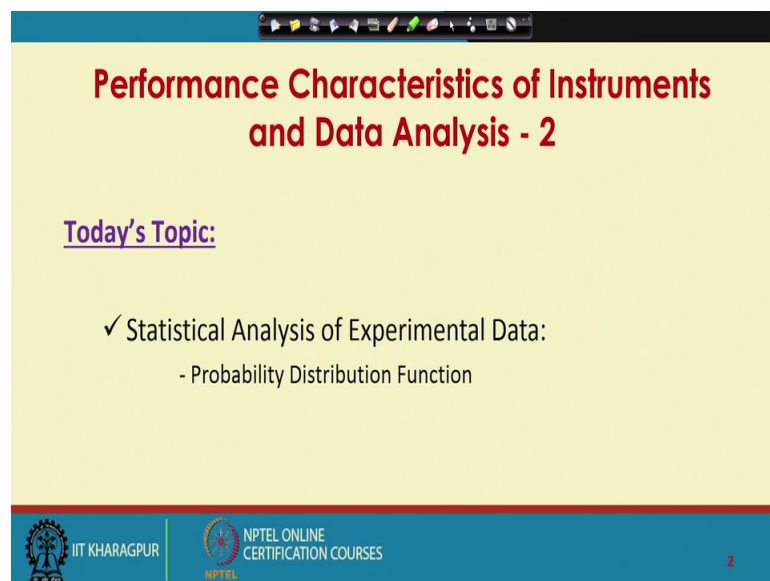


Chemical Process Instrumentation
Prof. Debasis Sarkar
Department of Chemical Engineering
Indian Institute of Technology, Kharagpur

Lecture – 14
Performance Characteristics of instruments and Data Analysis-II (Contd.)

Welcome to lecture 14, we have been talking about performance characteristics of instruments and data analysis part 2.

(Refer Slide Time: 00:23)



**Performance Characteristics of Instruments
and Data Analysis - 2**

Today's Topic:

✓ Statistical Analysis of Experimental Data:
- Probability Distribution Function



IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, today's topic is statistical analysis of experimental data and there which we will mainly focus on probability distribution functions and we also we will we will mainly talk about normal distribution function or Gaussian distribution functions.

(Refer Slide Time: 00:46)

Probability Distribution

- Any measurement is associated with several factors which cause random error.
- The instrument readings exhibit a dispersion/scatter of data.
- If a measurement is taken for a very large number of times, a probability distribution can be obtained from the data.



Any measurement is associated with several factors which cause random error the instrument readings exhibit a dispersion or scatter of data this is because the measurement as associated with several factors which cause random error. So, the measurement readings will always exhibit a dispersion or scatter of data.

If a measurement is taken for a very large number of times a probability distribution can be obtained from the data and that is what we learnt in this lecture.

(Refer Slide Time: 01:28)

Probability Distribution

60 Temperature Measurements

Number of readings	Temperature (°C)
1	1089
1	1092
2	1094
4	1095
8	1098
9	1100
12	1104
4	1105
5	1107
5	1108
4	1110
3	1112
2	1115

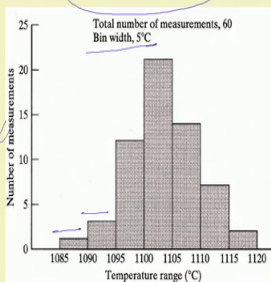
$$\bar{X} = \frac{x_1 + x_2 + x_3 + \dots + x_N}{N} = 1103 \text{ }^\circ\text{C}$$

Median = 1104 °C
Mode = 1104 °C

Standard deviation:
$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})^2} = 5.79 \text{ }^\circ\text{C}$$



Variance:
$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})^2 = 33.49 \text{ }^\circ\text{C}^2$$

Histogram plot



Total number of measurements, 60
Bin width, 5°C

How many readings fall in each bin?



Let us consider this table this table shows sixty temperature measurements of some hot object the readings are arranged in order of temperature. So, it ranges from 1089 to 1115 degree Celsius on the left. This is number of readings what it means is 1089 temperature 1089 degree Celsius was obtained, once 1094 was obtained twice 1104 was obtained 12 times 1108 was obtained 5 times so on and so forth. So, on the left, this number of readings essentially indicates the frequency of distribution or the frequency of the occurrence.

Now, given this data we can find out the basic statistical parameters such as mean the definition we have learnt in the previous lecture median can be obtained as 1104 degree Celsius mode is also 1104 degree Celsius, if you remember mode is the observation that occurs most frequently and this is what happens here 1104 has happened 12 times. So, mode is 1104 degree Celsius standard deviation can be computed using this formula where x_i is the individual observation \bar{x} is mean which is this and capital N is the number of observations this standard deviation happens to be 5.79 degree Celsius variance is square of standard deviation and that is obtained as 33.49 degree Celsius look at the formula of standard deviation.

We divide this quantity by $N - 1$ for a very large number of readings whereas, it will not matter whether you divide it by N or $N - 1$, but smaller size samples will always be divided by $N - 1$ accordingly, they are known as biased or unbiased estimates. So, we can calculate these statistical parameters easily now we can also have an histogram plot. So, by histogram plot what we first ask is if I divide this temperature range into small bins of say 0.5 degree Celsius each how many readings will fall in each bin. So, from 1085 to 1090, I have considered 1 bin 1090 to 1095, I have considered another bin so on and so forth, then from here and here I count how many readings are there in each bin. So, this is that number of measurements or readings. So, basically this is the frequency of occurrence.

So, this plot this is known as a histogram plot. So, it pictorially presents the distribution of the data.

(Refer Slide Time: 06:54)

Probability Distribution

Define: $Z = \frac{\text{No. of reading in an interval (bin)}}{\text{Total no. of readings} \times \text{Width of interval (bin)}}$

The area of a bar is equal to the probability that a specific reading will fall in the associated interval. The area of the entire histogram must then be 1.

If we have large number of readings and thus can make the bin width very small, the steps in the graph would become smaller and smaller, and the graph would approach a smooth continuous curve. The function $Z = f(x)$ is then called the probability density function. Here x is value of temperature (random variable)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

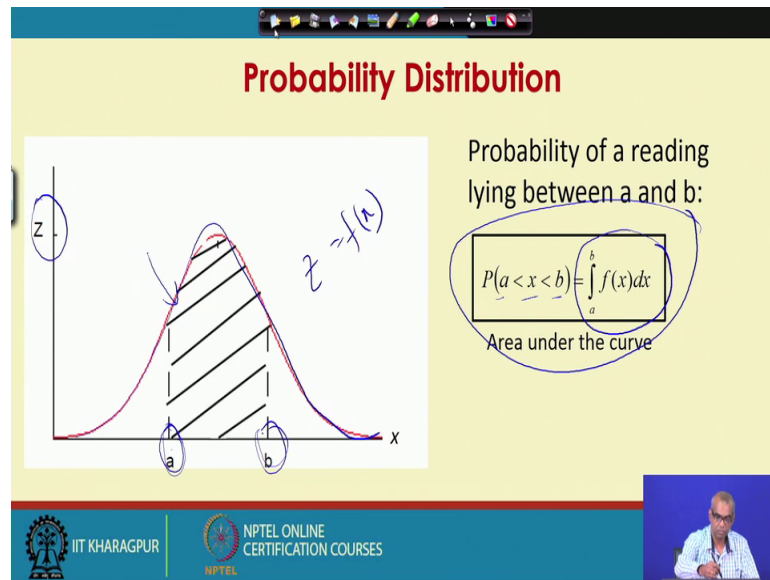
Now, let us define a term Z which is as follows I find out number of readings in each bin or interval whatever you call it divided by total number of readings and divided by width of each bin or interval. So, with this definition of Z , I now plot Z versus those temperature readings now the area of each bar the area of each bar is equal to the probability that a specific reading will fall in the associated interval. So, the area of this bar is the numerical value of the probability that a specific reading will fall in this interval. So, it is naturally follows that the area of the entire histogram must be one because a reading will have probability one to fall anywhere within the histogram.

If we have a very large number of readings and that is can make the bin width very very small the steps in the graph would become smaller and smaller and the graph would approach a smooth continuous curve what I mean is if I really have very very large number of readings, then I can have very small intervals or very small size bins and even then I can expect that the bins will have some finite number of readings in it under this situation the steps in the graph would becomes smaller and smaller and the graph would approach a smooth continuous curve in the limit of very large number of readings the function Z equal to $f(x)$ is then called the probability dens density function here x is the value of the temperature which is being considered as a random variable here.

So, what I mean is if I now consider these bins very very small, then perhaps you can have a continuous distribution like this, but before that you have this very very small

bins throughout and if you really have very large number of data we can do it if you have really very very large number of that data we can do this and still expect that it be each bin will have some finite number of readings in it. So, the function Z equal to $f(x)$ under such situation will be called the probability density function.

(Refer Slide Time: 11:11)



So, the probability of a reading lying between a and b is given by the area under the curve between these 2 points. So, this is Z equal to $f(x)$ relation you can consider it was initially an histogram and then I made the bin size is extremely small. So, that I was able to draw this smooth continuous curve.

So, what is the probability that a reading will lie between point a and b; that means, what is a probability that x will lie between a and b that can be obtained by finding out the area under the probability density function curve between points a and b. So, the area under the curve can be computed by this integral. So, the probability that x will lie between a and b is equal to integral a to b $f(x) dx$ you know that this term represent the area under the curve $f(x)$ here $f(x)$ is z .

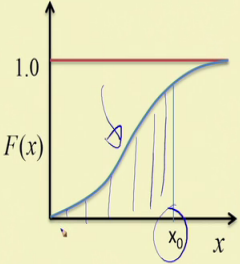
(Refer Slide Time: 12:57)

Probability Distribution


Probability information is also given by cumulative distribution.



Cumulative Distribution:
Probability that a reading is less than any chosen value of x :

$$F(x) = \int_{-\infty}^{x_0} f(x) dx$$



The cumulative distribution function is the area under the curve to the left of a vertical line drawn through point x_0 .



The probability information can also be given by cumulative distribution probability that a reading is less than any chosen value of x represents cumulative distribution. So, the cumulative distribution is defined by this integral. So, it is the probability that a reading is less than any chosen value of x . So, this is the cumulative distribution consider the point x_0 consider the value of the observation as x_0 .

So, what is the probability that a that a reading is less than x_0 for that we find the area under the curve to the left of a vertical line drawn through point x_0 . So, this area.

(Refer Slide Time: 14:30)

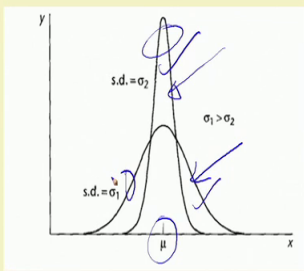
Probability Distribution: Gaussian Distribution

Gaussian (Normal) Distribution is the most important probability distribution in the statistical analysis of experimental data.


Most random errors follow Gaussian distribution.



$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / (2\sigma^2)}$$

μ - mean, σ - std dev



Normal distributions with same mean but different values of standard deviation

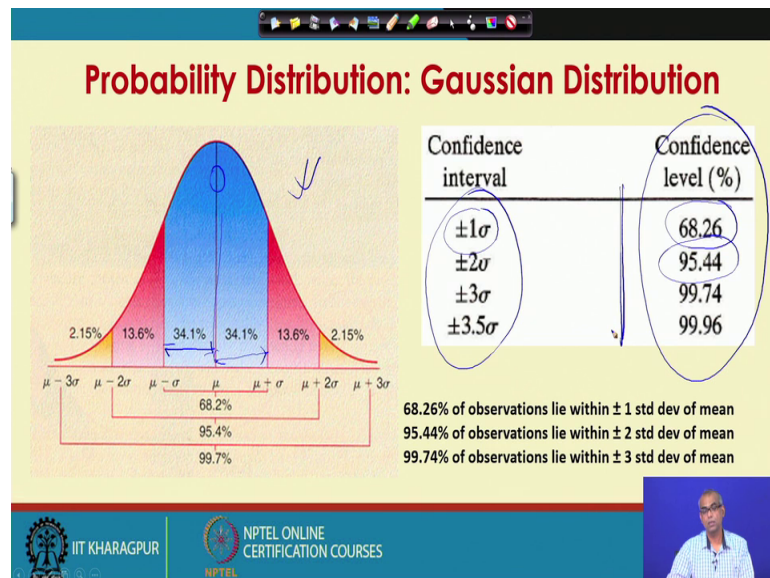


There are several probability distribution functions and among these Gaussian distributions is very popular for performing statistical analysis on experimental data the Gaussian distribution is also known as normal distribution most random errors follow Gaussian distribution. So, what is Gaussian distribution let us look at the expression that describes Gaussian distribution a Gaussian distribution is characterized by 2 terms; one is mean of the distribution another is standard deviation of the distribution and the Gaussian distribution is defined by this expression $P(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Where μ represents mean of the distribution and σ represents standard deviation. So, the Gaussian distribution or normal distribution looks like a bell shaped curve there are 2 distribution shown in this figure both has same mean μ , but different standard deviations this one has standard deviation σ_1 and these one has standard deviation σ_2 standard deviation σ_1 is greater than σ_2 in this figure standard deviation represents the spread of the distribution or scatter of the data. So, this one is a distribution where the scatter of data is less than this distribution.

(Refer Slide Time: 17:17)



Gaussian distribution has several interesting properties this represents Gaussian distribution and this represents the mean of the distribution.

It is known that if you consider an interval which is plus minus one standard deviation around mean, then 62.26 percent of the observations will fall within this interval. So,

what I mean is you have this mean of the distribution you consider μ plus sigma and μ minus sigma the Gaussian distribution is symmetric around the mean. So, in μ plus sigma in this interval 34.1 percent of all the observations will lie similarly in μ minus sigma another 34.1 percent will lie. So, within μ plus minus sigma sixty two point six percent will lie. So, we call plus minus one sigma as confidence interval and the corresponding confidence level is 62.26 percent.

Now, if you go to 2 sigma plus minus 2 sigma around the mean; that means, if I increase now the length of the interval by if I increase the length of the interval to 2 sigma around the mean it will be μ plus 2 sigma μ minus 2 sigma within this interval 95.44 percent of the observations will lie similarly ninety nine point seven four percent observations will lie within plus minus three standard deviation of the mean. So, these confidence intervals or these confidence levels can be computed for the normal distribution function remember this numbers this is very interesting property of normal distribution.

(Refer Slide Time: 20:24)

Probability Distribution: Confidence Interval

We wish to make an estimate of the population mean, which takes the form

$$\mu = \bar{x} \pm \delta \quad \text{or} \quad \bar{x} - \delta \leq \mu \leq \bar{x} + \delta$$

where δ is an uncertainty and \bar{x} is the sample mean. The interval from $\bar{x} - \delta$ to $\bar{x} + \delta$ is called the *confidence interval* on the mean. However, the confidence interval depends on a concept called the *confidence level*, sometimes called the *degree of confidence*. The confidence level is the probability that the population mean will fall within the specified interval:

$$\text{confidence level} = P(\bar{x} - \delta \leq \mu \leq \bar{x} + \delta)$$

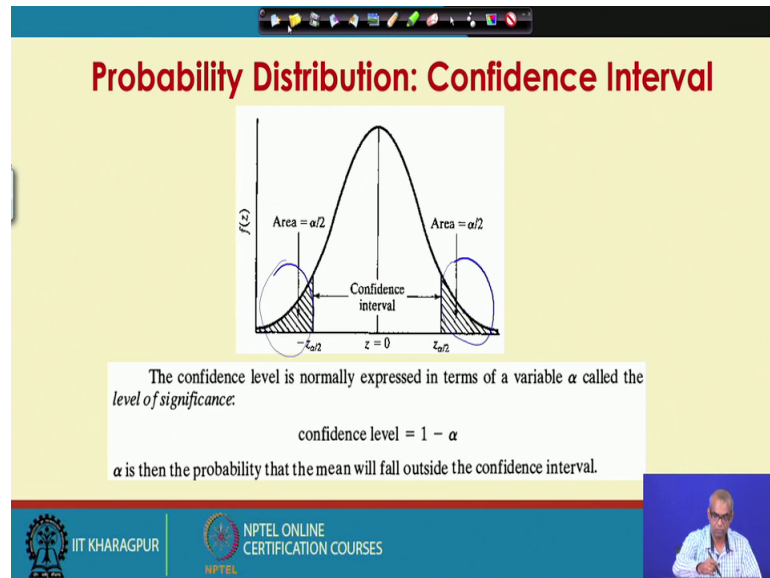
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Few more words about confidence interval we wish to make an estimate of the population mean which takes the form μ equal to \bar{x} plus minus delta. So, mean eq.

So, mean equal to \bar{x} plus minus delta. So, μ lies between \bar{x} minus delta to \bar{x} plus delta where delta is an uncertainty and \bar{x} is the sample mean the interval from \bar{x} minus delta to \bar{x} plus delta is called the confidence interval on the mean; however, the confidence interval depends on a concept called the confidence level

sometimes called the degree of confidence the confidence level is the probability that the population mean will fall within the specified interval. So, confidence level is the probability that the population mean will fall within the specified interval.

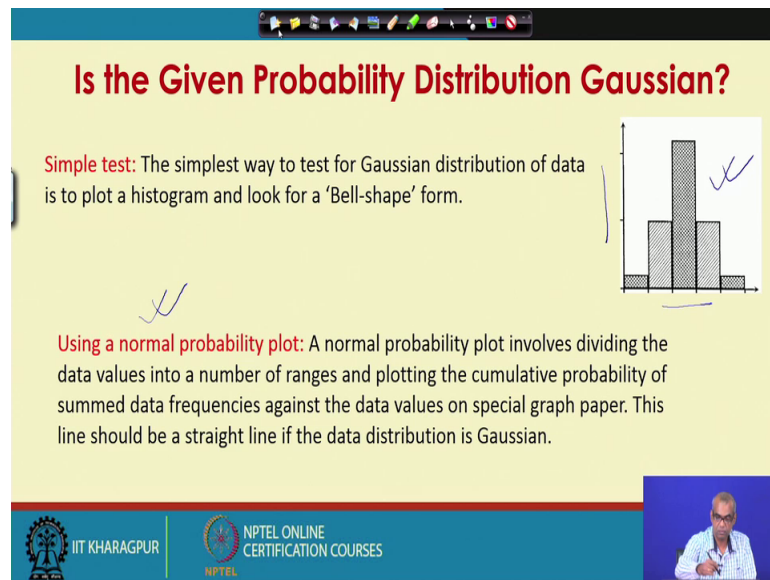
(Refer Slide Time: 21:43)



The confidence level is normally expressed in terms of a variable alpha called as level of significance and confidence level is one minus alpha. So, one minus level of significance.

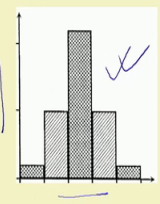
So, alpha is then probability that the mean will fall outside the confidence interval. So, if this area is alpha by 2 and this area is also alpha by two. So, alpha by 2 plus alpha by 2 is alpha. So, alpha is the probability that the mean will fall outside the confidence interval. So, the conf. So, the confidence level is 1 minus alpha.

(Refer Slide Time: 22:40)



Is the Given Probability Distribution Gaussian?

Simple test: The simplest way to test for Gaussian distribution of data is to plot a histogram and look for a 'Bell-shape' form.



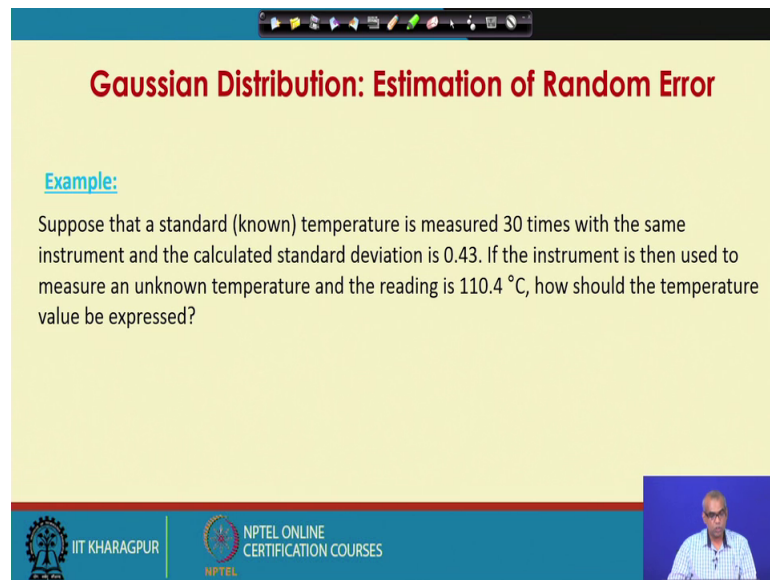
Using a normal probability plot: A normal probability plot involves dividing the data values into a number of ranges and plotting the cumulative probability of summed data frequencies against the data values on special graph paper. This line should be a straight line if the data distribution is Gaussian.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

How do I know that a given probability distribution is Gaussian or normal in nature you have done a series of experiments suppose you have measured a standard known temperature several times or you are measuring a random variable several times now you have generated a set of data now how do you know that this data will follow a normal or Gaussian distribution there are 2 quick test for this one is a simple test the simplest way to test for Gaussian distribution of data is to plot a histogram and look for a bell shape form.

So, these are the readings and these are the occurrence frequency of the occurrence. So, the histogram looks like bell shape you know that the distribution is Gaussian another is making use of normal probability plot a normal probability plot involves dividing the data values into a number of ranges and plotting the cumulative probability of some data frequencies against the data values on special graph paper. So, there is a special graph paper available and you have a normal probability plot there the line should be a straight line if the data distribution is Gaussian.

(Refer Slide Time: 24:34)



Gaussian Distribution: Estimation of Random Error

Example:

Suppose that a standard (known) temperature is measured 30 times with the same instrument and the calculated standard deviation is 0.43. If the instrument is then used to measure an unknown temperature and the reading is 110.4 °C, how should the temperature value be expressed?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Let us ask a simple question suppose that a standard or known temperature is measured thirty times with the same instrument and the calculated standard deviation is point four three if the instrument is then used to measure an unknown temperature and the reading is 110.4 degree Celsius how should the temperature value is expressed.

So, the question you ask is we know a standard temperature known standard temperature we measure it thirty times and calculate the standard deviation as 0.43; if the instrument is then used to measure an unknown temperature and the reading is 110.4 degree Celsius how should the temperature value is expressed; that means, I have to express the temperature as 110.4 degree Celsius plus minus the uncertainty associated with the reading. So, in the previous lecture we have talked about standard error which is standard deviation divided by square root of number of observation.

(Refer Slide Time: 25:56)

Gaussian Distribution: Estimation of Random Error

Solution:

Standard error of the mean, $U_n = \frac{\sigma}{\sqrt{n}} = \frac{0.43}{\sqrt{30}} = 0.08$

Let us calculate the error within 95% confidence limits, i.e. we calculate the value of the deviation D such that 95% of the area under the probability curve lies within limits of $\pm D$. This value is $\pm 1.96\sigma$.

Thus, the maximum likely error in a single measurement can be expressed as:
Error = $\pm(1.96\sigma + U_n) = \pm((1.96)(0.43) + 0.08) = \pm 0.92$.

Thus, express unknown temperature as : $110.4 \pm 0.9^\circ\text{C}$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, let us first calculate the standard error of the mean. So, standard error of the mean can be calculated as standard deviation divided by square root of observation number of observations 30 here.

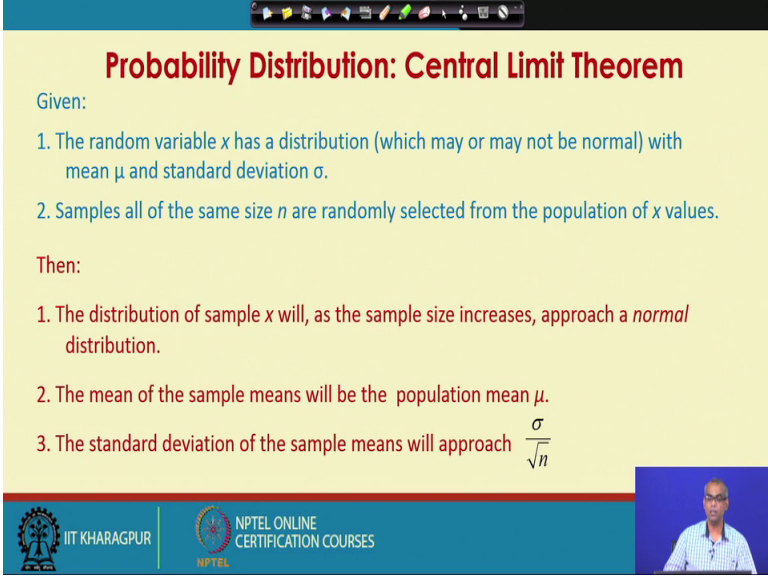
So, I calculate the standard error as 0.08; let us now calculate the error within 95 percent confidence limits that is we calculate the value of the deviation D such that 95 percent of the area under the probability curve lies within the plus minus within limits of plus minus D. So, this is something which you must understand. So, we wish to calculate the error within 95 percent confidence limit. So, you have to first associate a confidence limits. So, let us say that we want to calculate the error within 95 percent confidence limits what it means is that you want to calculate the value of the deviation D such that 95 percent of the area under the probability curve will lie within limits of plus minus D, it is like one sigma within plus minus 1 sigma you have 68.2 percent readings will lie.

Similarly, within plus minus 2 sigma 95.44 percent of all the readings will lie. So, you ask the question within plus minus what sigma say x. So, within plus minus x sigma 95 percent of the readings will lie. So, what is the value of x if it is 2 sigma is 95.44, we intuitively feel that x will be very close to 2 actually it is 1.96. So, within plus minus 1.96 sigma 95 percent of the observations will lie. So, you get this value as 1.96 into sigma. So, the maximum error the maximum likely error in a single measurement can be now expressed as error equal to plus minus 1.96 sigma plus the standard error which is

standard deviation by square root of N. So, you compute this sigma value you put as point four three this standard error also obtained as 0.08.

So, error is plus minus 0.92. So, the unknown temperature can now be expressed as 110.4 plus minus 0.9 degree Celsius.

(Refer Slide Time: 29:18)



Probability Distribution: Central Limit Theorem

Given:

1. The random variable x has a distribution (which may or may not be normal) with mean μ and standard deviation σ .
2. Samples all of the same size n are randomly selected from the population of x values.

Then:

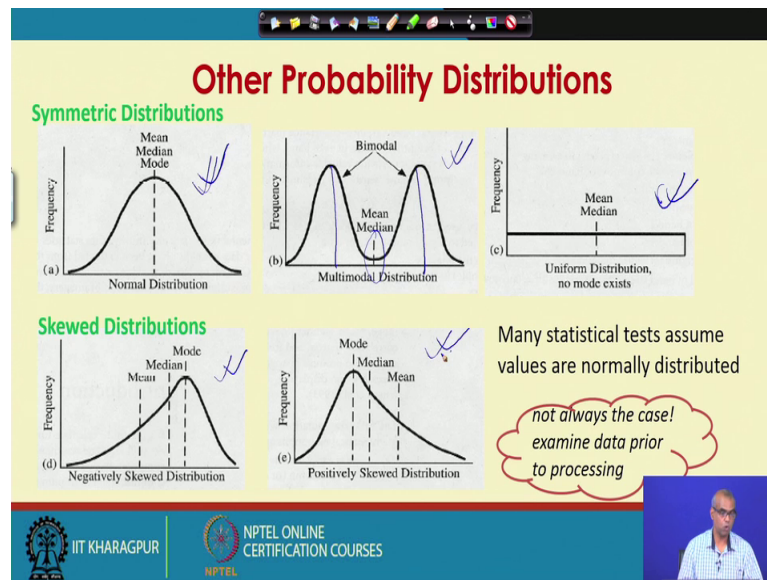
1. The distribution of sample x will, as the sample size increases, approach a *normal* distribution.
2. The mean of the sample means will be the population mean μ .
3. The standard deviation of the sample means will approach $\frac{\sigma}{\sqrt{n}}$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Let us learn about the theorem known as central limit theorem given the random variable x has a distribution which may or may not be normal with mean μ and standard deviation σ samples all of the same size n are randomly selected from the population of x values, then the distribution of sample x will as the sample size increases approach a normal distribution the mean of the sample means will be the population mean μ the standard deviation of the sample means will approach σ by square root of n . So, what it says is if you have a random variable x which has a distribution with mean μ and standard deviation σ .

And samples of samples all of the same size n are randomly selected from the population of x values then the distribution of sample x will approach a normal distribution as the sample size increases the mean of the sample means will be the population mean μ the standard deviation of the sample means will approach standard deviation divided by square root of n which is n equal to number of observations.

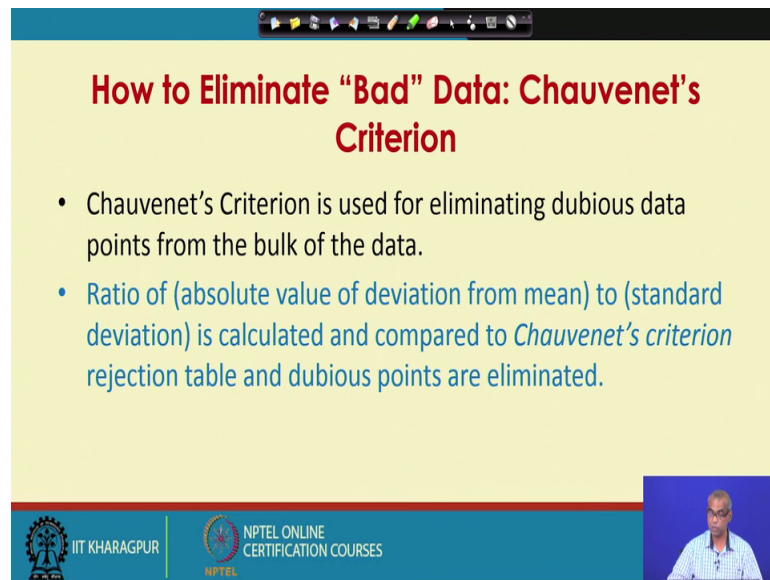
(Refer Slide Time: 30:54)



There are other probability distributions symmetric distributions where mean equal to median equal to mode you have a single mode here you have 2 modes here this is also symmetric. So, you have one mode here you have another mode here and mean and median is this this is uniform distribution.

These are non symmetric or asymmetric distributions this is negatively skewed this is positively skewed note that in case of non symmetric or skewed distributions mean median mode are all different many statistical test assume values are normally distributed, but be careful this is not always the case always examine data prior to processing. So, when you are measuring a particular quantity and you have a set of data and then you see that one or 2 data do not follow the trend of the remaining data. So, you ask yourself what was wrong in performing experiments. So, is there an way to eliminate such bad data Chauvenet's criterion serves this purpose it helps you to eliminate the bad data.

(Refer Slide Time: 32:51)



How to Eliminate "Bad" Data: Chauvenet's Criterion

- Chauvenet's Criterion is used for eliminating dubious data points from the bulk of the data.
- Ratio of (absolute value of deviation from mean) to (standard deviation) is calculated and compared to *Chauvenet's criterion* rejection table and dubious points are eliminated.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Which are outlier Chauvenet's criterion is used for eliminating dubious data points from the bulk of the data ratio of absolute value of deviation from mean to standard deviation is calculated and compared to Chauvenet's criterion rejection table and dubious points are eliminated. So, what you do have to do is once you have the set of observation or set of data you find out the absolute value of deviation from mean and divide that by the standard deviation for each reading and then compare it to the Chauvenet's criterion rejection table which is available and then the dubious points can be eliminated by comparing these 2 data new mean and standard deviations can be computed after you eliminate the outliers or dubious data points.

(Refer Slide Time: 33:56)

Number of readings, n	Ratio of maximum acceptable deviation to standard deviation
3	1.38
4	1.54
5	1.65
6	1.73
7	1.80
10	1.96
15	2.13

Chauvenet's criterion for rejecting a "bad" data

So, this is an example of Chauvenet's rejection table.

So, this is ratio of maximum acceptable deviation to standard deviation and these are the number of readings. So, if you find that for a particular case your number is more than this that ratio of the ratio that you we just learned in the previous slide that absolute deviation from the mean divided by the standard deviation sees more than this for observation you can consider that to be a bad data and may be eliminated. So, we stop lecture 14 here.