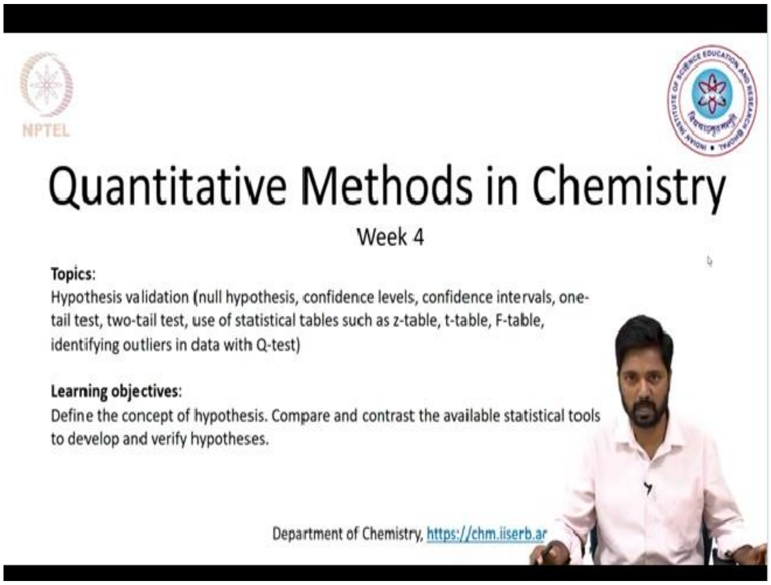
# Quantitative Methods in Chemistry

**Prof. Dr. Aasheesh Srivastava, Dr. Bharathwaj Sathyamoorthy**
**Department of Chemistry**
**Indian Institute of Science Education and Research-Bhopal**

## Lecture-13
## Introductory Statistics Part 01

Hello and welcome to this NPTEL course titled quantitative methods in chemistry. My name is Aasheesh Srivastava and I will be the instructor for this course for the next few weeks. So, this is week 4 of the course.

**(Refer Slide Time: 00:46)**



Quantitative Methods in Chemistry

Week 4

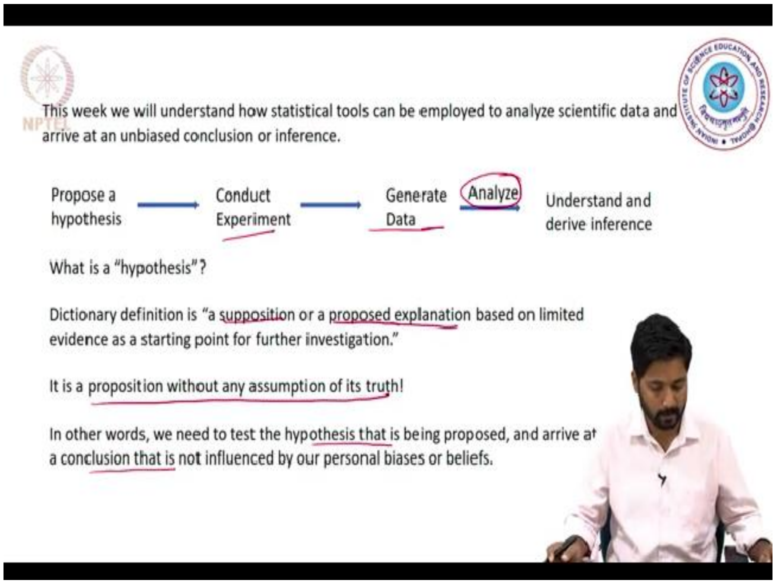
**Topics:**
Hypothesis validation (null hypothesis, confidence levels, confidence intervals, one-tail test, two-tail test, use of statistical tables such as z-table, t-table, F-table, identifying outliers in data with Q-test)

**Learning objectives:**
Define the concept of hypothesis. Compare and contrast the available statistical tools to develop and verify hypotheses.

Department of Chemistry, https://chm.iiserb.ac

And this week we will be learning about hypothesis validation, what is null hypothesis, what are confidence levels, confidence intervals, what are two-tail test, one-tail test and use of statistical tables such as z-table t-table, F-table etc. for analyzing scientific data. So, the objectives for this week's course are to define the concept of hypothesis and compare and contrast the available statistical tools to develop and verify various hypothesis.

**(Refer Slide Time: 01:29)**



This week we will understand how statistical tools can be employed to analyze scientific data and arrive at an unbiased conclusion or inference.

Propose a hypothesis → Conduct Experiment → Generate Data → Analyze → Understand and derive inference

What is a “hypothesis”?

Dictionary definition is “a supposition or a proposed explanation based on limited evidence as a starting point for further investigation.”

It is a proposition without any assumption of its truth!

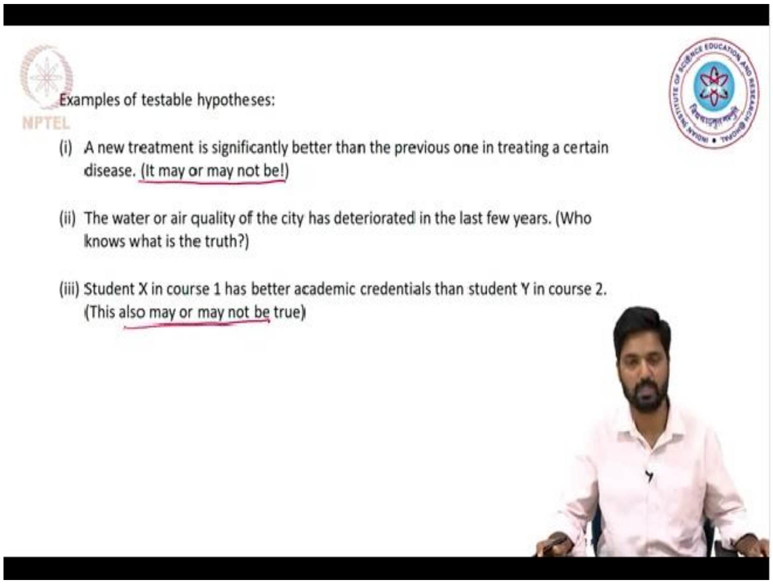
In other words, we need to test the hypothesis that is being proposed, and arrive at a conclusion that is not influenced by our personal biases or beliefs.

So, let us begin this week's classes. So, this week we will understand how statistical tools can be employed to analyze scientific data and arrive derive pattern unbiased conclusion. Usually, we will have a hypothesis that we will propose and then we will conduct experiments to test our hypothesis. These experiments will generate a set of data and when we analyze this data through a variety of tools, we understand what this data is telling us.

And then we derive at a inference from what this data is telling us. However, before we move further, it is important to delve in what the term hypothesis itself means, if we go by the dictionary meaning it is a supposition or a proposed explanation based on our limited evidence, which acts as a starting point for further investigation, it is also important to emphasize that it is a proposition without any assumption of its truth.

So, a hypothesis needs to be tested for its truthfulness. And we need to test this hypothesis and arrive at a conclusion that is not influenced by our personal believes and biases.

**(Refer Slide Time: 03:03)**



Examples of testable hypotheses:

(i) A new treatment is significantly better than the previous one in treating a certain disease. (It may or may not be!)

(ii) The water or air quality of the city has deteriorated in the last few years. (Who knows what is the truth?)

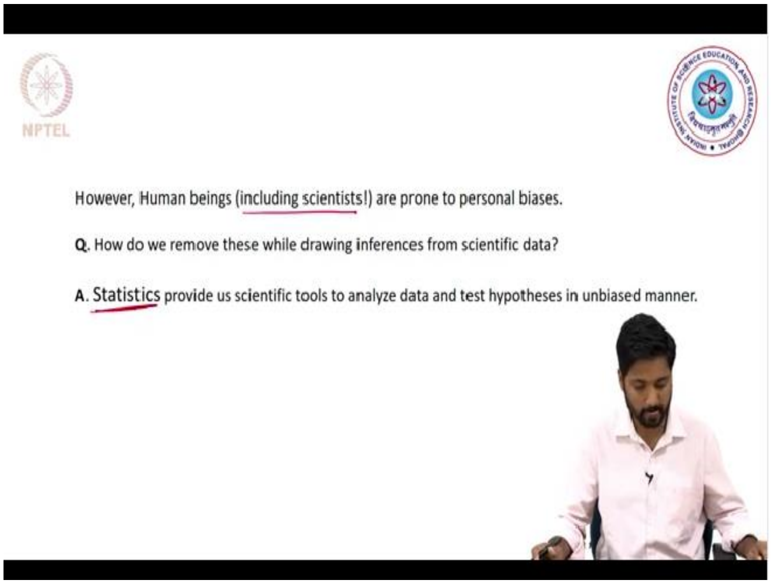
(iii) Student X in course 1 has better academic credentials than student Y in course 2. (This also may or may not be true)

And how do we do this, before that let us see some example of hypotheses that are testable or that we can test through generation of scientific data. For example, let us presume that we created a treatment for a certain illness. Now we need to test whether this treatment is actually working on the disease that we are utilizing this treatment for. To test that we will propose the hypothesis that the new treatment is significantly better than the previous one in treating the disease.

However, as is written here, this hypothesis may or may not be true, and we need to test this hypothesis for its truthfulness. Another example is the water or the air quality of the city has deteriorated in the last few years. Now, we do not know whether this is true or false. And again, this is a hypothesis which is testable through generation of scientific data. Another example of the same concept is let us take a student x, who is enrolled in course 1.

And we want to test whether the student x has better academy credentials than the student y who is enrolled in a different course, which we call as course 2. So again this the fact that student x is having better academy credentials than student y may or may not be true, and this again becomes a testable hypothesis.

**(Refer Slide Time: 04:49)**



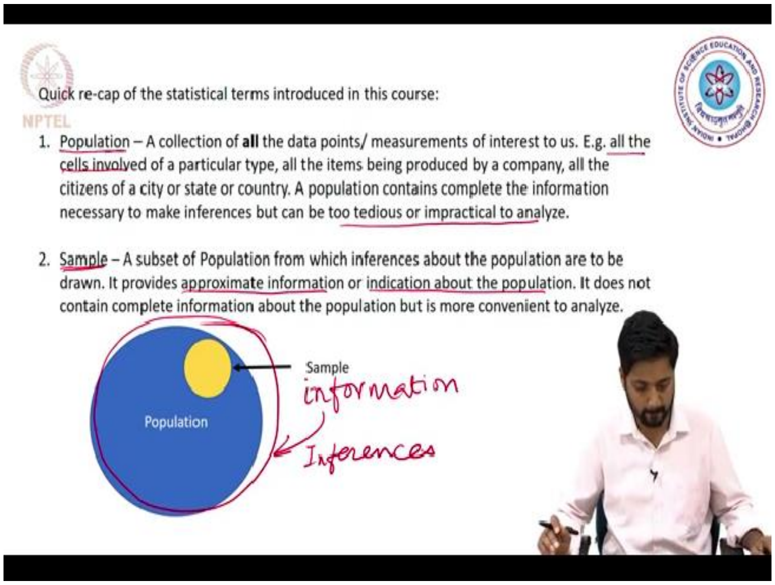
However, Human beings (including scientists!) are prone to personal biases.

Q. How do we remove these while drawing inferences from scientific data?

A. Statistics provide us scientific tools to analyze data and test hypotheses in unbiased manner.

Now, when we are testing our hypothesis, we as human beings, this includes researchers and scientists like us, we are prone to our own personal biases. And the question that we have in our hand today is how do we remove our biases while drawing inferences from scientific data. And the answer to that is we employ statistics for doing this, statistics provide us with scientific tools to analyze the data and test our hypothesis in an unbiased manner.

**(Refer Slide Time: 05:28)**



Quick re-cap of the statistical terms introduced in this course:

1. Population – A collection of **all** the data points/ measurements of interest to us. E.g. all the cells involved of a particular type, all the items being produced by a company, all the citizens of a city or state or country. A population contains complete the information necessary to make inferences but can be too tedious or impractical to analyze.

2. Sample – A subset of Population from which inferences about the population are to be drawn. It provides approximate information or indication about the population. It does not contain complete information about the population but is more convenient to analyze.

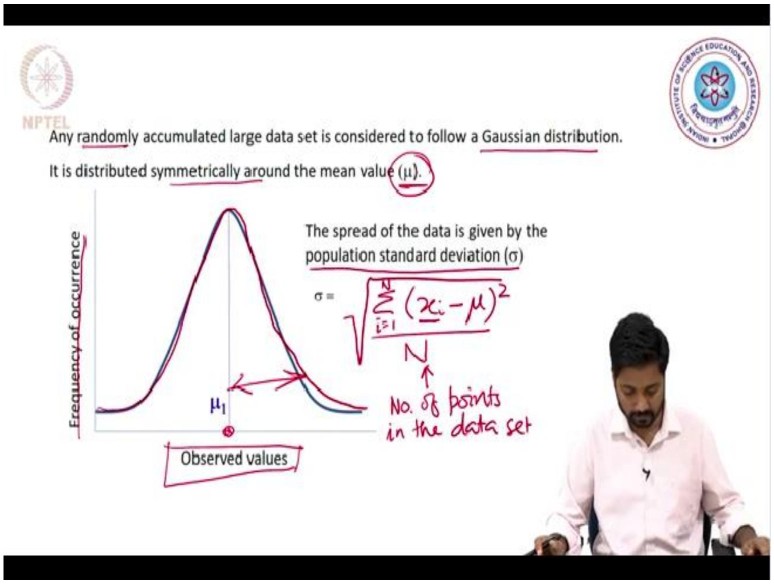
So, before we delve further, let us take a very quick recap of what the statistical terms that have been introduced in this course. One of the terms that you might have encountered is the concept of population, in statistics population is the collection of all the data points or measurements that are of interest to us. And when we say all we mean actually all. For example, when we are talking about progression of disease, we say that all the cells that are involved in a particular type of disease.

Similarly, it can be all the items that particular companies producing or all the citizens of a country or a city or a state. And in nutshell, a population contains complete information necessary to make inferences that however, when we start analyzing the whole population, this can be too tedious or often impractical to analyze. So, what is the solution. The solution is that we take small subset from the population and make our inferences for the whole population.

Now, this subset is known as the sample. So, to rephrase a sample is a subset of the population from which we will derive inferences about the population. It provides an approximate information or an indication about the population that we are dealing with, it does not contain the complete information about the population. However, analyzing smaller sample is more convenient. Now, the same concept can be understood by use of the Venn diagrams.

For example, this circle which we have drawn here indicates set of data points which will call as the population. Now, as I told before, analyzing this whole population may be too tedious. So, we will utilize a small subset from this population which we will call the sample and we would utilize the information contained in this sample to make inferences about the whole population.

**(Refer Slide Time: 08:21)**



Any randomly accumulated large data set is considered to follow a Gaussian distribution.

It is distributed symmetrically around the mean value (μ).

The spread of the data is given by the population standard deviation (σ)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

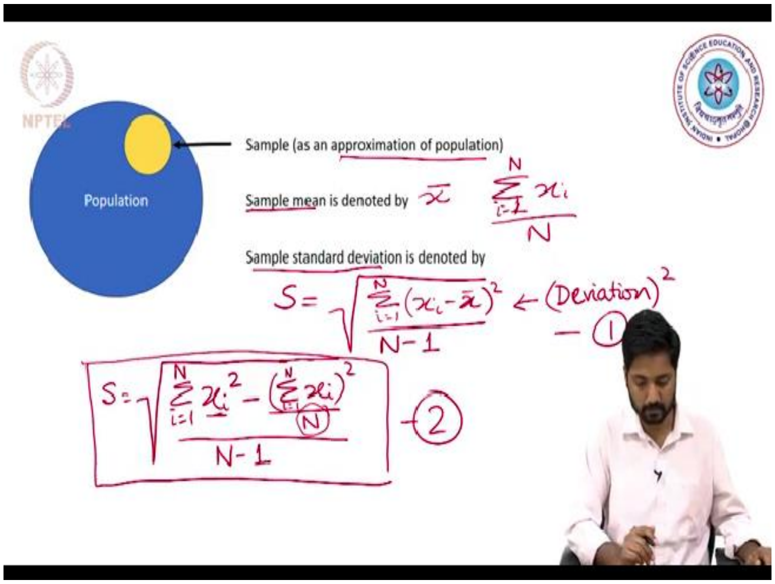
No. of points in the data set

Another point that has been emphasized in this course is that when we collect large set of data and we collected in a random manner, that means we do not put our own influences or biases into it, this data will follow a Gaussian distribution. So, it would be distributed symmetrically around the main value which we indicated by mu. For example, this profile which we have drawn here is what is known as a Gaussian profile.

Now, as we can see very clearly, the data points are distributed quite evenly across the central point, which is the mean value of this data set. So, when we plot the frequency of occurrence versus the observed values we get this Gaussian profile, from this Gaussian profile, it becomes easy to analyze the data through mathematical formula. For example, we can easily figure out what is the mean value of this data set.

Similarly, the spread of this data set is given by what is known as the population standard deviation or sigma. Now the sigma is given by mathematical formula which looks something like this. So, xi is the individual values of this data set, mu is the mean value of the data set and n is the number of points in the data set. So, if we have the information about mu and sigma, we can very well understand how this population of data is behaving.

**(Refer Slide Time: 11:01)**



Sample (as an approximation of population)

Sample mean is denoted by $\bar{x}$ $\quad \frac{\sum_{i=1}^{N} x_i}{N}$

Sample standard deviation is denoted by

$$S = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1}} \leftarrow (\text{Deviation})^2 \quad - (1)$$

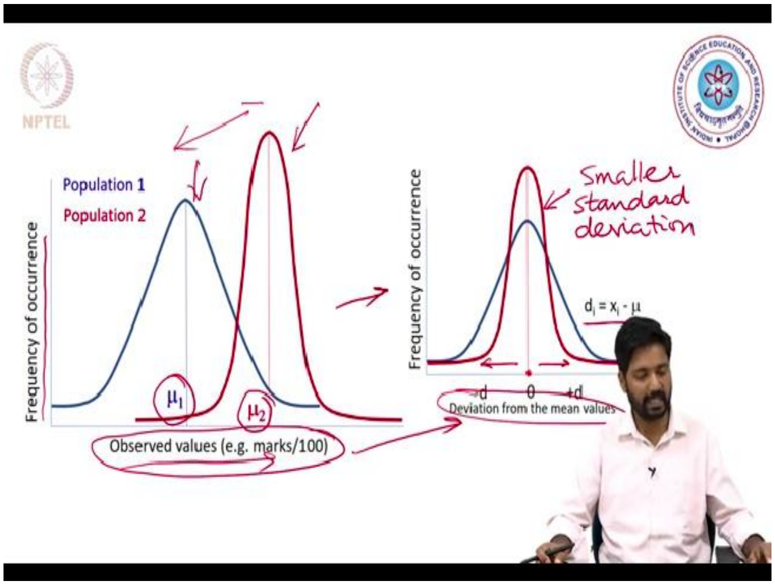
$$\boxed{S = \sqrt{\frac{\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}}{N-1}}} \quad - (2)$$

Now, we have also told that from a population, we can take small sample, which is considered as an approximation of the population if it is taken in a random fashion. And, this sample can also be analyzed using statistics to arrive at similar concepts of sample mean, which now we denote by x bar, which again is summation of all the individual values divided by the number of values present in the sample.

Similarly, we can have a sample standard deviation now, which is denoted by S and is given by the formula written on the board, so here the numerator is nothing but what we call as the deviation squared. And n - 1 value comes from the fact that for small samples, we use the degrees of freedom. So essentially, the sample standard deviation boils down to the square root of the deviation squared divided by the degrees of freedom.

This can also be written as the formula shown here and both equation 1 as well as equation 2 will essentially give you the same values of the sample standard deviation. Now, let us look at the equation 2 a little bit more in detail, equation 2 tells us that we need to take the individual values which are squared and submit all of them and subtract from this summation the individual values some together and then whole square. And that divided by the number of observations n and then this whole thing is being divided by the degrees of freedom okay.

**(Refer Slide Time: 14:34)**



Now, let us consider a simple example where we are dealing with 2 different populations. Now, these populations are shown in this slide as the blue profile and the red profile. So, population 1 is denoted in blue while the population 2 is denoted in red. And we have plotted in this graph, the frequency of occurrence and the observed values. For example, let us take that these 2 populations are those of students who are taking 2 different courses okay.

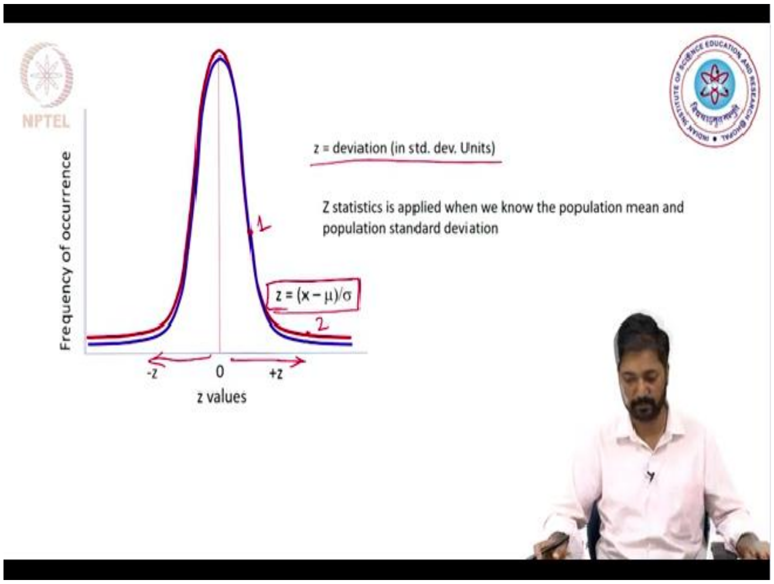
So, in these 2 courses the students have got different marks, their marks distribution is different and even the mean values of the marks obtained for the 2 courses are different, for course 1 the mean value is given by mu 1 and for course 2 the mean value is given by mu 2. Now, the question is how do we analyze these disjoint set of data and try to arrive at a conclusion whether a particular point is distinctly different from the other.

Now, let me delve on a little bit further. So, what we will do here is if we look at the observed values profile, these population 1 and population 2 are well separated. Now, how do we bring them together to analyze them together. One simple way to do this is to convert the observed values into the deviation from the mean values. And this deviation is given by the formula di, which is equal to the x i – mu.

And x i are the individual values and mu is the population mean. So, when we do that the blue and the red profiles are populations now merged around the central point which is the point at which the deviation value is 0. Now the deviation can be negative as well as positive and which is what is clearly shown in this profile. Now, even here we have a slight challenge in the fact that the red profile seems to have smaller standard deviation compared to the blue profile.

So, it may still not be possible for us to analyze these 2 data together. One further step will help us merge these 2 data together for an easier analysis.

**(Refer Slide Time: 17:57)**



z = deviation (in std. dev. Units)

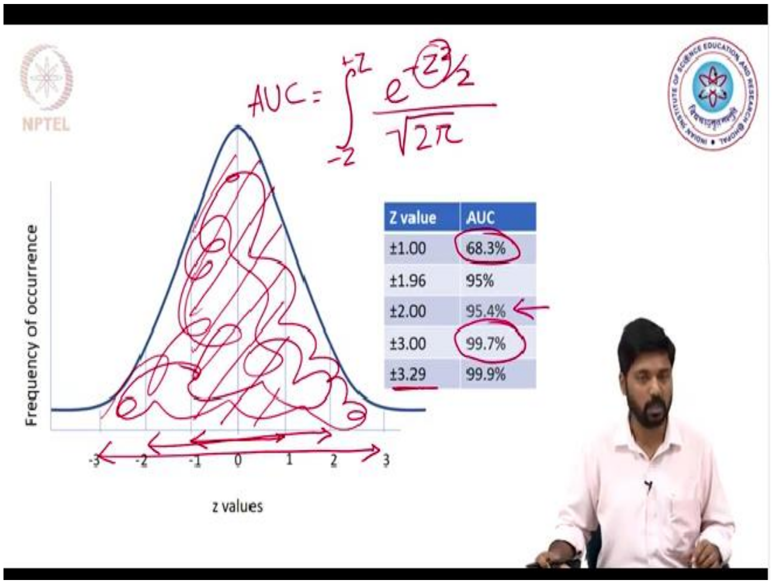
Z statistics is applied when we know the population mean and population standard deviation

$z = (x - \mu)/\sigma$

And that is done in the next slide where we have used what is known as a z statistics, which is the deviation values divided by the standard deviation. So, in other words, the z statistics is nothing but the deviation values in standard deviation units. Again, the values can be z values can be positive as well as negative. But in these z profile what we see is that the blue and red data now merge quite completely over one another.

So, if you want to compare point 1 versus on the blue curve versus point 2 on the red curve, it would be easier for us to analyze these 2 distinct or disjoint set of data.

**(Refer Slide Time: 19:00)**



$$AUC = \int_{-z}^{+z} \frac{e^{-z^2/2}}{\sqrt{2\pi}}$$

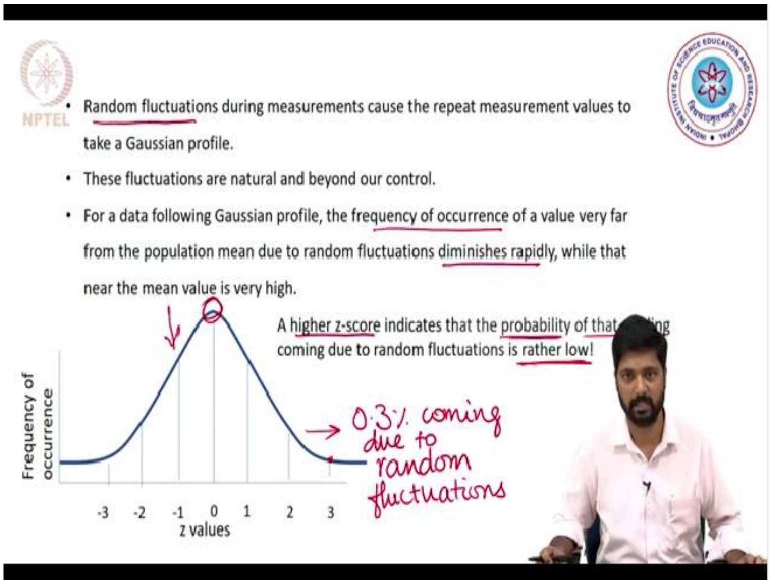
| Z value | AUC |
|---|---|
| ±1.00 | 68.3% |
| ±1.96 | 95% |
| ±2.00 | 95.4% |
| ±3.00 | 99.7% |
| ±3.29 | 99.9% |

So when we convert the Gaussian error function in terms of z, it takes the form which is shown here. So you integrate it from - z to + z to get the area under the curve, and it takes the form of e to the power of - z square by 2 by square root 2 pi. So, if we integrate this we get the area under the curve. Now when we do that what we observe is that, based on the z value that we plug in this equation, we get a variety of the area under the curve values.

So, for example, if I take the z value from – 1 to + 1, then we get the area under the curve being 68.3%. So, which for example we can shade like this. So, this area which has been shaded now is 68.3% of all the data points that we are dealing with. Similarly, when we expand this 2 + - 2 z values, then we are able to cover a larger area under the curve and this corresponds to 95.4% of all the data points that we are dealing with.

As we go further, we can make it to z + - 3, where we are now able to cover a very large proportion of the area which is correspondent to 99.7% of all the values in order to cover 99.9% of all the values, we would have to use a z value of + - 3.29. So, as we can see that this profile is very similar to that to what we obtain when we plot the error function with respect to the sigma value or the population standard deviation.

**(Refer Slide Time: 21:31)**



- Random fluctuations during measurements cause the repeat measurement values to take a Gaussian profile.
- These fluctuations are natural and beyond our control.
- For a data following Gaussian profile, the frequency of occurrence of a value very far from the population mean due to random fluctuations diminishes rapidly, while that near the mean value is very high.

A higher z-score indicates that the probability of that ng coming due to random fluctuations is rather low!

Now a few points to recollect here are that this Gaussian error function is coming due to random fluctuations during our measurements and when we repeat our measurements, if we do that in an unbiased manner or the measurement is supposed to take Gaussian profile which is demonstrated here. Now, these fluctuations in our measurements are considered to be natural and beyond our control.

And for a Gaussian profile, the frequency of occurrence of a value which is very far from the mean value is very low. So, which is again mentioned here, it diminishes very rapidly as the value that we are measuring moves very far away from the mean value which is denoted here. Now, this has an important implication in the terms of the z score. So, what we replace this statement as is if we have a higher z score that indicates that the probability of that reading or that measurement.

So, probability needs to be highlighted coming due random fluctuations is very, very low, which again we can see in this profile. So, if I have a z value of say, + 3, that it is 99.7% not likely to be coming due to the random fluctuations. So, the probability of this point with a high z score of + 3 is only 0.3% coming due to random fluctuations in our measurement. That means this data point with us z score of + 3 is considered to be statistically quite significant.