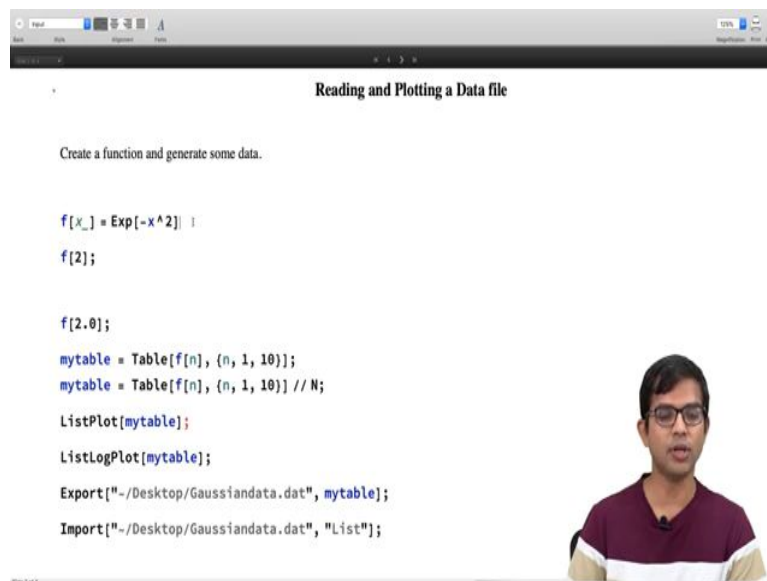


Physics through Computational Thinking
Doctor Auditya Sharma & Doctor Ambar Jain
Department of Physics
Indian Institute of Science Education and Research, Bhopal
Lecture 16
Introduction to Data Analysis 1

Hello, so in this lecture we're going to look at some aspects of data analysis and how this can be realized using Mathematica. Then we learn a few tricks of Mathematica along the way but also learn some very useful and important concepts of, how to compute error bars and oftentimes we have to deal with data of many kinds you know it could be experimental data, it could be simulation data, it could be you know data that you do get it from some some source let us say, and so there there are some basic statistical ideas that we should all be familiar with, and hopefully this lecture will shed light on on how to carry out this kind of data analysis.

(Refer Slide Time: 1:15)



```
Reading and Plotting a Data file

Create a function and generate some data.

f[x_] = Exp[-x^2] ;
f[2];

f[2.0];
mytable = Table[f[n], {n, 1, 10}];
mytable = Table[f[n], {n, 1, 10}] // N;

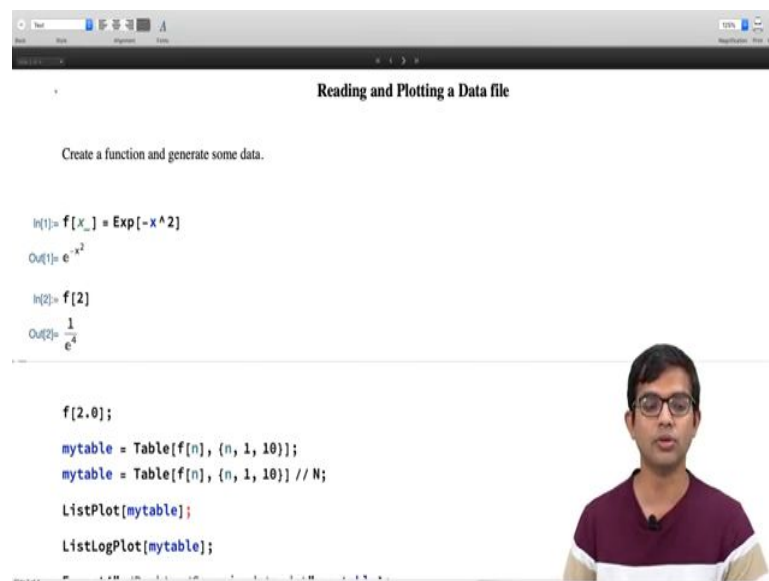
ListPlot[mytable];
ListLogPlot[mytable];

Export["~/Desktop/Gaussiandata.dat", mytable];
Import["~/Desktop/Gaussiandata.dat", "List"];
```

Okay, so first we will look at some aspects of Mathematica, right? So, we know that often we would like to plot a function for example right? So, with the programming language one of the key tasks that we would want to do is to define a function and to be able to plot it. In Mathematica the syntax for a function is like here, for example if I am interested in looking at the function e^{-x^2} , Gaussian.

I would write it like this, you know the capital E is important for exponential, so you this is something that you can easily look up and so when you define a function, so there is syntax with an x underscore, that you need to look at. So, maybe in a separate video or elsewhere we will go into some details of you know various different kinds of functional forms.

(Refer Slide Time: 2:27)



The screenshot shows a Mathematica notebook window with the title "Reading and Plotting a Data file". The notebook content includes the following text and code:

Create a function and generate some data.

```
In[1]:= f[x_] = Exp[-x^2]
```

Out[1]= e^{-x^2}

```
In[2]:= f[2]
```

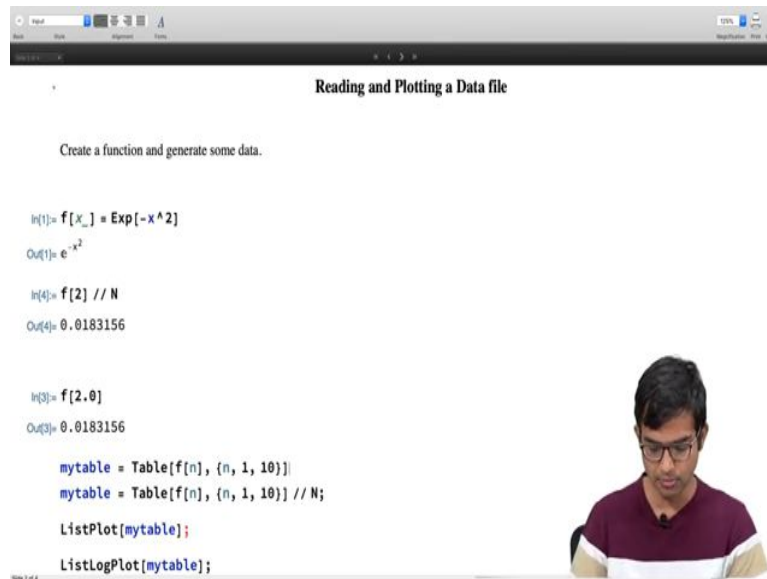
Out[2]= $\frac{1}{e^4}$

```
f[2.0];  
mytable = Table[f[n], {n, 1, 10}];  
mytable = Table[f[n], {n, 1, 10}] // N;  
ListPlot[mytable];  
ListLogPlot[mytable];
```

A small video inset in the bottom right corner shows a man with glasses and a maroon shirt speaking.

So, for now just accept that $f[x_]$ comes in, and so the way to go ahead and start implementing this is to press shift and enter, so if I do shift enter it shows that I have defined something called e to the power $-x^2$. So, now of course I want to check whether Mathematica really understands this, so then I might just for a test, I could look at you know some particular argument, I can put in $f[2]$ and if I do shift enter it gives me $1/e^4$ right.

(Refer Slide Time: 2:50)



Bust so, if am interested in finding out the numerical value of this I could put in 2.0 and so that is one way of getting the numerical value but I could also do something like, $f[2] // N$, so this is completely equivalent I could go ahead and do it and you see that the same number 0.0183156 shows up, at time you can go ahead and play with this. So, there are various ways of doing this.

So, the N can come before the functional evaluation and you have to be careful about the syntax or it can come as a post processing operation, so which is given by these kinds of two slanting lines which are leaning forward followed by N, that is a common standard approach.

So, then now suppose I want to create a whole table of these numbers, it is not for just one particular value, I want to study this function for a whole range of values, let us say I want I want to go from 1 to 10 in steps of one, then I can go ahead and use this syntax of a table, I can define a table, very useful function to know in Mathematica is that of table.

If you are familiar with other programming languages, you might have come across say a For loop in C or C++ and there are some advantages of using a For loop but oftentimes one of the powers of Mathematica is this table type of function which sometimes allows you to program for certain specific task in a in a much more efficient way alright, so this is a useful function to become aware of and to play with.

(Refer Slide Time: 4:39)

```

In[1]:= f[x_] = Exp[-x^2]
Out[1]= e^-x^2

In[4]:= f[2] // N
Out[4]= 0.0183156

In[3]:= f[2.0]
Out[3]= 0.0183156

In[7]:= mytable = Table[f[n], {n, 1, 10}]
mytable = Table[f[n], {n, 1, 10}] // N
Out[7]= {1/e, 1/e^4, 1/e^9, 1/e^16, 1/e^25, 1/e^36, 1/e^49, 1/e^64, 1/e^81, 1/e^100}
Out[8]= {0.367879, 0.0183156, 0.00012341, 1.12535 x 10^-7, 1.38879 x 10^-11,
2.31952 x 10^-16, 5.24289 x 10^-22, 1.60381 x 10^-28, 6.63968 x 10^-36, 3.72008 x 10^-44}

ListPlot[mytable] ;

```

So, now if I go ahead and hit shift enter, then so once again it just gives me these numbers with $1/e$, $1/e^4$ and so on. Mathematica is unwilling to give you numbers unless you force it to do so and the way to do that like we have seen before is to just put this slash N, so if I were to do this, then I get out all these numbers in gory detail. So, you can look up the documentation, may be there is going to be, like a separate video we might do to explain these kind of aspects, but yeah.

So, this is also something that you pick up as you go along, right? And once you have a table of such numbers, it is often of interest to get a plot of this kind of a list. And so the way to do that is to use a list plot.

(Refer Slide Time: 5:25)

```

In[3]:= f[2.0]
Out[3]= 0.0183156

In[7]:= mytable = Table[f[n], {n, 1, 10}]
mytable = Table[f[n], {n, 1, 10}] // N
Out[7]= { $\frac{1}{e}$ ,  $\frac{1}{e^4}$ ,  $\frac{1}{e^9}$ ,  $\frac{1}{e^{16}}$ ,  $\frac{1}{e^{25}}$ ,  $\frac{1}{e^{36}}$ ,  $\frac{1}{e^{49}}$ ,  $\frac{1}{e^{64}}$ ,  $\frac{1}{e^{81}}$ ,  $\frac{1}{e^{100}}$ }

Out[8]= {0.367879, 0.0183156, 0.00012341, 1.12535  $\times 10^{-7}$ , 1.38879  $\times 10^{-11}$ ,
2.31952  $\times 10^{-16}$ , 5.24289  $\times 10^{-22}$ , 1.60381  $\times 10^{-28}$ , 6.63968  $\times 10^{-36}$ , 3.72008  $\times 10^{-44}}$ 

In[9]:= ListPlot[mytable]

```

```

2.31952  $\times 10^{-16}$ , 5.24289  $\times 10^{-22}$ , 1.60381  $\times 10^{-28}$ , 6.63968  $\times 10^{-36}$ , 3.72008  $\times 10^{-44}$ 

In[9]:= ListPlot[mytable]

Out[9]= {0.367879, 0.0183156, 0.00012341, 1.12535  $\times 10^{-7}$ , 1.38879  $\times 10^{-11}$ ,
2.31952  $\times 10^{-16}$ , 5.24289  $\times 10^{-22}$ , 1.60381  $\times 10^{-28}$ , 6.63968  $\times 10^{-36}$ , 3.72008  $\times 10^{-44}}$ 

In[10]:= ListLogPlot[mytable]

Out[10]= {0.367879, 0.0183156, 0.00012341, 1.12535  $\times 10^{-7}$ , 1.38879  $\times 10^{-11}$ ,
2.31952  $\times 10^{-16}$ , 5.24289  $\times 10^{-22}$ , 1.60381  $\times 10^{-28}$ , 6.63968  $\times 10^{-36}$ , 3.72008  $\times 10^{-44}}$ 

In[11]:= Export["~/Desktop/Gaussiandata.dat", mytable];
In[12]:= Import["~/Desktop/Gaussiandata.dat", "List"];

```

And if I were to do this, so you see that for large values of N, so it is practically 0. So it is basically on the x axis beyond a certain point. And it is not a surprise, of course, you can go ahead and beautify you know, these plots using x labels for the x axis and y axis and so on. That is all something that you can look up by looking up the documentation of list plot.

So, there is also list log plot, which sometimes is more useful, particularly in in you know, cases like here, when the data falls off so rapidly with x that you are hardly able to notice any difference for larger N and but if you plot it on a log plot, so it is log along the y axis and linear along the the x axis. So, then you see that, you know, it is exaggerated, the fall is

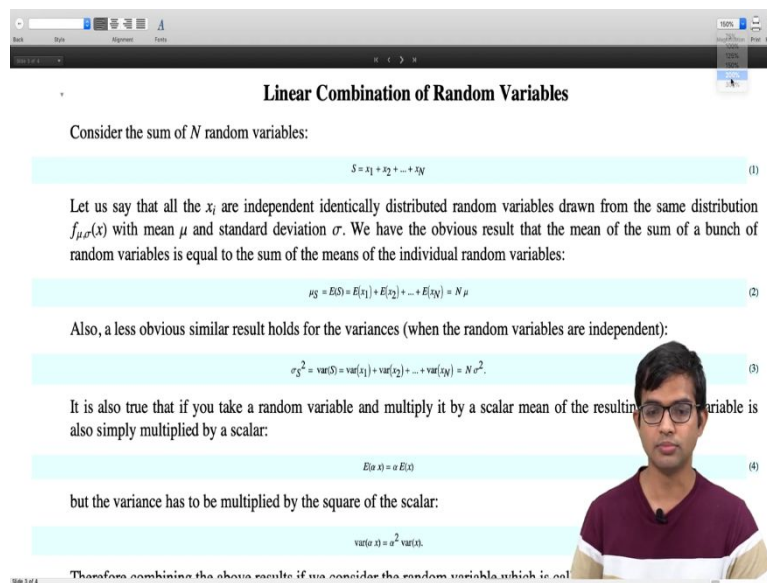
exaggerated because of the logarithmic variation along the y axis. So, this is a useful command as well.

So, once you have figured out how to generate such data and to visualize data, sometimes it is useful to be able to just store this data. So, one way to do that is to use this function called export. So, you can go ahead and export an entire table to go to the desktop, you can put it wherever to put it on a folder of your choice on your machine. And then you can go ahead and export it in a dat file and there are advantages of just exporting it into a CSV file, so we might talk about it later on.

And so if you can export then you should also be able to import, you can also use this complementary function called import. So, you have to make sure that you provide the correct path to the file where it is located on your computer. And so you can import it and then put it up in, onto Mathematica so that if you want to carry out analysis.

So, this is the kind of thing that would happen. For example, say if you have carried out an experiment and you have experimental data, and it could be in a CSV file or a dat file in like here, and then you you want to bring it on your machine and only then we would be going ahead with carrying out the kind of analysis we will describe now. So, this is some basics of the syntax of Mathematica for carrying out the data analysis.

(Refer Slide Time: 8:00)



Linear Combination of Random Variables

Consider the sum of N random variables:

$$S = x_1 + x_2 + \dots + x_N \quad (1)$$

Let us say that all the x_i are independent identically distributed random variables drawn from the same distribution $f_{\mu, \sigma}(x)$ with mean μ and standard deviation σ . We have the obvious result that the mean of the sum of a bunch of random variables is equal to the sum of the means of the individual random variables:

$$\mu_S = E(S) = E(x_1) + E(x_2) + \dots + E(x_N) = N \mu \quad (2)$$

Also, a less obvious similar result holds for the variances (when the random variables are independent):

$$\sigma_S^2 = \text{var}(S) = \text{var}(x_1) + \text{var}(x_2) + \dots + \text{var}(x_N) = N \sigma^2 \quad (3)$$

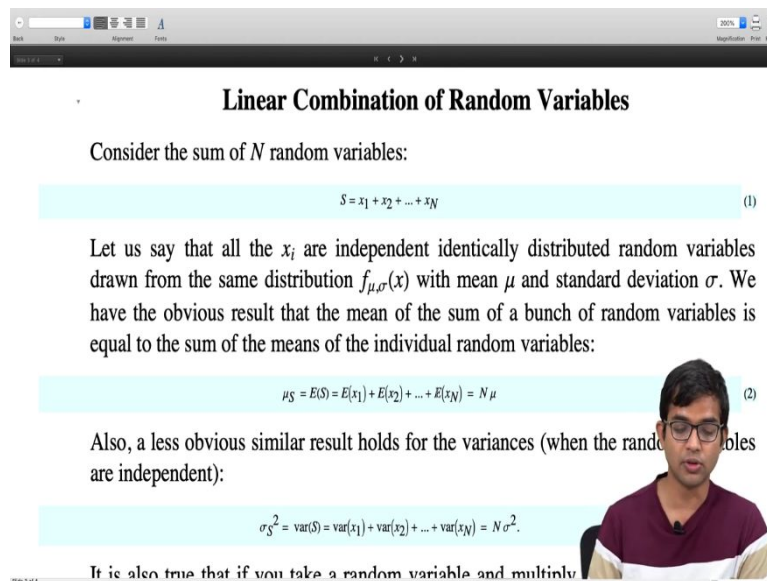
It is also true that if you take a random variable and multiply it by a scalar mean of the resulting variable is also simply multiplied by a scalar:

$$E(a x) = a E(x) \quad (4)$$

but the variance has to be multiplied by the square of the scalar:

$$\text{var}(a x) = a^2 \text{var}(x).$$

Therefore combining the above results if we consider the random variable which is equal to the sum of N independent identically distributed random variables...



Linear Combination of Random Variables

Consider the sum of N random variables:

$$S = x_1 + x_2 + \dots + x_N \quad (1)$$

Let us say that all the x_i are independent identically distributed random variables drawn from the same distribution $f_{\mu, \sigma}(x)$ with mean μ and standard deviation σ . We have the obvious result that the mean of the sum of a bunch of random variables is equal to the sum of the means of the individual random variables:

$$\mu_S = E(S) = E(x_1) + E(x_2) + \dots + E(x_N) = N \mu \quad (2)$$

Also, a less obvious similar result holds for the variances (when the random variables are independent):

$$\sigma_S^2 = \text{var}(S) = \text{var}(x_1) + \text{var}(x_2) + \dots + \text{var}(x_N) = N \sigma^2 \quad (3)$$

It is also true that if you take a random variable and multiply it by a scalar mean of the resulting variable is also simply multiplied by a scalar: $E(a x) = a E(x)$ but the variance has to be multiplied by the square of the scalar: $\text{var}(a x) = a^2 \text{var}(x)$.

Therefore combining the above results if we consider the random variable which is equal to the sum of N independent identically distributed random variables...

Okay, so now we will go into some theory of data analysis. So, some of you must have seen random variables, you might have taken a course on elementary probability theory. So, even if you have not, so the idea here is not to give you any sort of rigorous introduction into these ideas, but hopefully, the ideas here will be sort of intuitively clear.

And so from this discussion, we want to take home a prescription. So, this is a very important prescription. And so, we will sort of delineate the key ideas which go into, you know, how one finds an estimate for the mean and how one finds an estimate for the error bar, and so on.

So, there's a lot of theory on this. So if you want all the details, and if you want to figure out a rigorous approach to this, then that that would perhaps be a course in itself. But for now, here, we want to just pick out the essence. So, the essence here is the following. So, let us say you have N random variables. So, and we are interested in looking at the sum of these N random variables.

So imagine doing an experiment, and then taking N readings right? So, you are measuring some distance, you are measuring some length or time, or whatever. But let us say that you carry out this experiment N different times. And so you are going to get N different numbers. So they are all perhaps going to be in the same ballpark, but they are not all going to be identical, right? So that is the whole point of you know, a random variable is something that comes from a distribution.

Now, let us imagine that these random variables are independent and identically distributed. So, this is a sort of a basic assumption that one makes to model these kinds of random variables. I mean, of course, there are other random variables which may not be totally independent, you know, there may be some variation in distribution and all that but at the simplest level, we model them as being independent and identically distributed. Identically distributed means that all of them come from the same distribution and independent means that the value of you know one measurement does not influence another. So, each of them is completely independent.

Now, let us say that they are drawn from a distribution $f_{\mu,\sigma}(x)$, and these μ and σ are the parameters which characterize your distribution. So, there is a mean μ and a standard deviation σ of course, you can have even higher moments.

So, the whole distribution of course, contains information about all the moments if you know everything about all the moments, then you basically know the distribution itself but μ , mean and standard deviation are like the the most rudimentary pieces of information about the distribution, they carry some information, but if you want to know everything, then you will have to know all the moments or the entire distribution itself.

So, given the distribution, you can compute the mean, you can compute the standard deviation. But if you want to compute the whole distribution, you need not just the first two

moments, but you need all the moments. So, this is the theory of probability, which you might encounter in a different course. But Ok so let us go through this in an intuitive way, we have the obvious result that if you take the mean of the sum right, I said that we are often interested in calculating the sum of this kind of the distribution of x itself.

And so one thing that is obvious is that if you take the sum and find its average, find the expectation value of the sum, it is going to be just the sum of the expectation values of each of these individual random variables. And since each of them is identically distributed, each of them has the the mean μ , it is going to be just $\mu + \mu + \mu + \dots$, so on N times, and that just gives you $N * \mu$.

Ok. So, this was straightforward enough, but also a less obvious result holds for the variances right? So, this is something that you can work out. Maybe we can post this as a homework problem. So, you can work out the result that if you were to add the variances of, you know all these x_1 and x_2 and x , all the way up to x_N also add to give you the variance of the overall random variable itself.

So, just like the sum of the expectation values is equal to the expectation value of the sum. So, likewise the sum of the variances is equal to variance of the sum. So, that is a result which I am going to just state and I, it is something that you can, you can show very quickly, very quickly, you can prove it with just applying the definition of variance.

(Refer Slide Time: 13:26)

$S = x_1 + x_2 + \dots + x_N$ (1)

Let us say that all the x_i are independent identically distributed random variables drawn from the same distribution $f_{\mu, \sigma}(x)$ with mean μ and standard deviation σ . We have the obvious result that the mean of the sum of a bunch of random variables is equal to the sum of the means of the individual random variables:


$\mu_S = E(S) = E(x_1) + E(x_2) + \dots + E(x_N) = N\mu$ (2)

Also, a less obvious similar result holds for the variances (when the random variables are independent):

$\sigma_S^2 = \text{var}(S) = \text{var}(x_1) + \text{var}(x_2) + \dots + \text{var}(x_N) = N\sigma^2$ (3)

It is also true that if you take a random variable and multiply it by a scalar, the mean of the resulting random variable is also simply multiplied by a scalar:

$E(\alpha x) = \alpha E(x)$



So, the variance is simply defined as the second moment right, so given the distribution, you know how to compute the mean is going to be $\int x f dx$ and the average of x^2 is $\int x^2 f dx$. And then you can go ahead and compute the mean of x^2 . And the mean of x is already known.

And if you take the difference of these two, you will get the variance. And so you can go ahead and put in this definition and work out this relation. So, anyway the key point is that you get σ_s^2 is equal to $N * \sigma^2$. So, each of these has the same variance because it is a identically distributed, and therefore we have this result σ_s^2 equal $N * \sigma^2$.

So, it is also true that if you take a random variable and multiply it by a scalar, the resulting random variables simply multiply by a scalar. So, if you take expectation value of some scalar times x , it is going to be α times the scalar will come out and expectation value will remains as it is, but if you have if you are doing something similar with variances, variance of $\alpha * x$ is going to be $\alpha^2 * \text{Var}(x)$ right. So, now it appears as α^2 this is important.

(Refer Slide Time: 14:22)

but the variance has to be multiplied by the square of the scalar:

$$\text{var}(\alpha x) = \alpha^2 \text{var}(x). \quad (5)$$

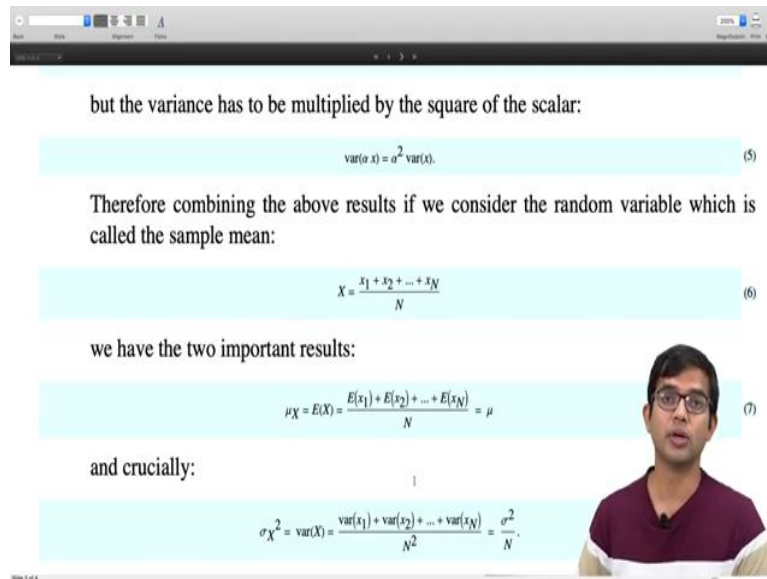
Therefore combining the above results if we consider the random variable which is called the sample mean:

$$X = \frac{x_1 + x_2 + \dots + x_N}{N} \quad (6)$$

we have the two important results:

$$\mu_X = E(X) = \frac{E(x_1) + E(x_2) + \dots + E(x_N)}{N} = \mu \quad (7)$$

and crucially:

$$\sigma_X^2 = \text{var}(X) = \frac{\text{var}(x_1) + \text{var}(x_2) + \dots + \text{var}(x_N)}{N^2} = \frac{\sigma^2}{N}$$


And so a consequence of this is, suppose you are looking at not just the sum of these random variables right. So, this is sort of an intermediate step, we are interested mainly in the mean of these random variables. Suppose, you are looking at x is equal to the sum of all these random variables divided by N , then you can go ahead and convince yourself that the mean of all of this is simply equal to the mean of the distribution, right.

So, because you have each of these $\langle x_1 \rangle + \langle x_2 \rangle$ so on, and then you have this factor of N which comes in the denominator, which will just simply cancel with the end that you had in the numerator, and therefore μ_X is nothing but its just μ .

And crucially, so if you were to do σ_X^2 , so now you have this important α^2 term right. So, when you, when you take this variance of each of these guys, we will go we will go with an α outside, and that α is just $1/N$. And then when you pull it out, you get $1/N^2$.

And then there is the N , which is sitting in the numerator, because you have $N * \sigma^2$ in the numerator, which comes from all of these random variables being independent and identically distributed with variances, σ^2 , and then when you divide by N^2 , so the final answer is just $\sigma_X^2 = \sigma^2/N$ right.

(Refer Slide Time: 15:50)

we have the two important results:

$$\mu_X = E(X) = \frac{E(x_1) + E(x_2) + \dots + E(x_N)}{N} = \mu \quad (7)$$

and crucially:

$$\sigma_X^2 = \text{var}(X) = \frac{\text{var}(x_1) + \text{var}(x_2) + \dots + \text{var}(x_N)}{N^2} = \frac{\sigma^2}{N} \quad (8)$$

Often times when we perform experiments we acquire data whose distribution is unknown, so the precise μ and σ are not available. However the sample mean \bar{x} turns out to be an excellent estimate of the true mean of the distribution. The uncertainty in our estimate of the actual mean of the quantity is contained in the standard deviation $\sigma_{\bar{x}}$, which we have seen is given by $\frac{\sigma}{\sqrt{N}}$. Unfortunately the σ of the distribution itself is unknown, and with some (beautiful, not very difficult) arguments one can show that the sample variance:

which we have seen is given by $\frac{\sigma^2}{N}$. Unfortunately the σ of the distribution itself is unknown, and with some (beautiful, not very difficult) arguments, one can show that the sample variance:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_N - \bar{x})^2}{N - 1} \quad (9)$$

can yield an excellent estimate of the variance of the actual distribution with the aid of the simple relation:

$$\sigma^2 = \frac{N}{N-1} s^2 \quad (10)$$

The overall consequence of all this is that given a sample of N data points x_i , the mean along with error bars is given by the compact relation:

$$X \pm \frac{s}{\sqrt{(N-1)}} \quad (11)$$

So, think of these various numbers x_1, x_2, x_3 as different values you get for performing when you perform an experiment multiple times, and then you are interested in getting an estimate of the actual mean of the distribution. So, oftentimes we do not know what this μ is, or σ is for the, for the distribution from which it comes from.

And we want to be able to use experimental methods to determine to get an estimate of μ and σ . So, what we do is, so it turns out that you can just take an average of all these numbers x_1, x_2, x_3 , so on and you have seen that $\mu_{\bar{x}}$ is equal to μ . So, in fact the sample mean itself is a good estimate of the mean of the distribution.

So, that is the first result, you can just simply go ahead and take an average of a bunch of numbers that have come out of the experiment. And that already gives you an estimate of the mean of the distribution. And so if suppose we had known σ , then we could have computed σ_x^2 .

σ_x is what is a measure of the error in x ultimately if μ is an estimate of the mean, then σ_x is going to be the error in μ . So, we want to get at σ_x and to get at σ_x , we would need to know σ , because we have this relation $\sigma_x^2 = \sigma^2/N$. And if you know σ , then you can go ahead and immediately get σ_x . And if you have σ_x , then that gives you the sort of the spread about the mean about μ .

But the problem is that you do not already know what is σ . And so it turns out that actually there is a way to get an estimate of σ itself. So, we will not go into these arguments. So, they have some, you know, nice arguments which are available, if you are interested, you can look up a more advanced text on linear on data analysis.

But so believe me that there is a way to argue that this σ^2 is in fact related to the sample variance. So, when I say sample variance, I just mean, you know, you have these numbers x_1 , to x_N , using just these N numbers, you can extract a sample mean, which is just the mean of these numbers.

But you can also extract a standard deviation or a variance of the sample. And so that is simply given by the mean of the deviation squared right. So that is one way of thinking of variance right, x is the mean. So, you go to every random variable and subtract the mean from it, and square it. So, it is the deviation square. And if you take the mean of these guys, that gives you the sample variance.

So, I mean, so this overall, σ^2 is something like x^2 , but it is not exactly this. So, there are some arguments which go into this. And it turns out that the exact relation is σ^2 is $N/(N - 1) * s^2$. So, as you can see, if N is large, this factor $N/(N - 1)$ does not really matter so much you can, for all practical purposes, just take $\sigma^2 = s^2$.

But the key point is this: σ_x^2 is related to σ^2 with this factor of $1/N$. So that's a very important factor, maybe this $N/(N - 1)$ business is some detail. And oftentimes, it is of no consequence.

If your capital N is large, which is often the case you want to repeat an experiment many many, many times, get a lot of numbers and find a mean and as to estimate the mean of the distribution, and also use the sample variance to get an estimate of the error.

So, the overall consequence of all of this is that the sample mean gives us an estimate of the mean of the distribution but very, very crucially, it is the sample standard deviation divided by $\sqrt{(N-1)}$, which gives us an estimate of the error bars. So, this $\sqrt{(N-1)}$ which is sitting in the denominator is of great importance.

So, what it tells you is, by increasing the number of trials, if you do the experiment many, many more times, your error bars are going to shrink. If we did not have this $\sqrt{(N-1)}$ in the denominator, then it would mean that no matter how many trials you did, your data is not going to improve, but that is not the case. So, if you have a systematic result, so then you are going to have, so the more the trials, you have, the better the data, the better is the quality of the data and so, so the error bars are going to shrink. And so this is the relation which allows this.

And, oftentimes, this is something that is not not very widely known. This division by $\sqrt{(N-1)}$ and this is the kind of analysis which is carried out, say, when you do Monte Carlo simulations right. Some data may come from experimental situations, you have actually measured it in the lab or it could be data which has come out from computer experiments right. And so we might also discuss some of these Monte Carlo methods at a later time.

But yeah so just keep this point in mind. So, the mean of the sample mean itself is a good estimate of the actual mean of the distribution, which is, which is unknown, and the sample standard deviation $\sigma / \sqrt{(N-1)}$ to get an estimate of the error in your estimate of the mean. So, these are the two sort of take home messages.