# Data Analysis and Decision Making - I

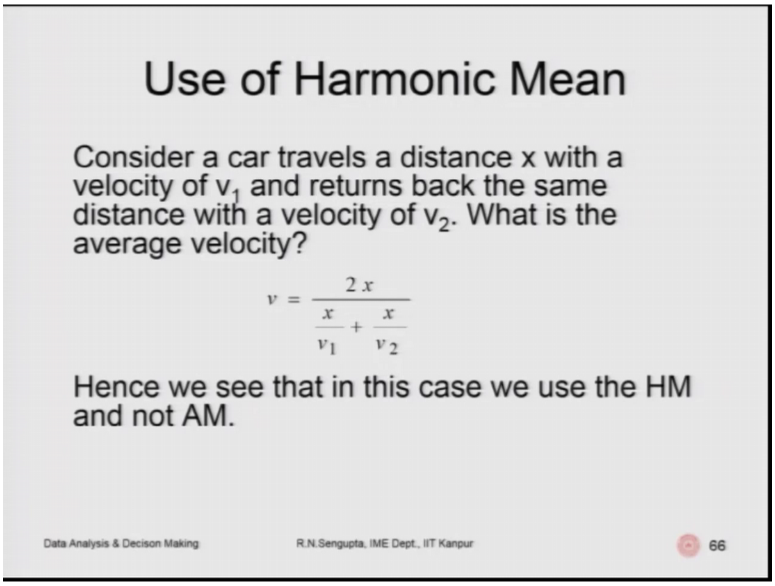
**Prof. Raghu Nandan Sengupta**
**Department of Industrial & Management Engineering**
**Indian Institute of Technology, Kanpur**

## Lecture – 04
## Descriptive Statistics

Good morning, good afternoon, good evening all my dear friends and welcome to this NPTEL MOOC lecture series in DA DM, which is Data Analysis and Decision Making which is the first of this 3 different courses, which is being plotted in NPTEL MOOC. Then this is the fourth lecture for DA DM I, which is for 30 hours.

So, we are in the first week and I am Raghu Nandan Sengupta from IME Department, IIT Kanpur. So, we are discussing in the examples of arithmetic mean and geometric mean and for the geometric mean I gave very simple example that how we can use the concept of interest rate to find out the overall average interest rate. So, let us continue the discussions further on using for further examples for harmonic mean and so on and so forth. Consider the example of the Harmonic mean.

(Refer Slide Time: 01:06)



Use of Harmonic Mean

Consider a car travels a distance x with a velocity of $v_1$ and returns back the same distance with a velocity of $v_2$. What is the average velocity?

$$v = \frac{2x}{\frac{x}{v_1} + \frac{x}{v_2}}$$

Hence we see that in this case we use the HM and not AM.

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 66

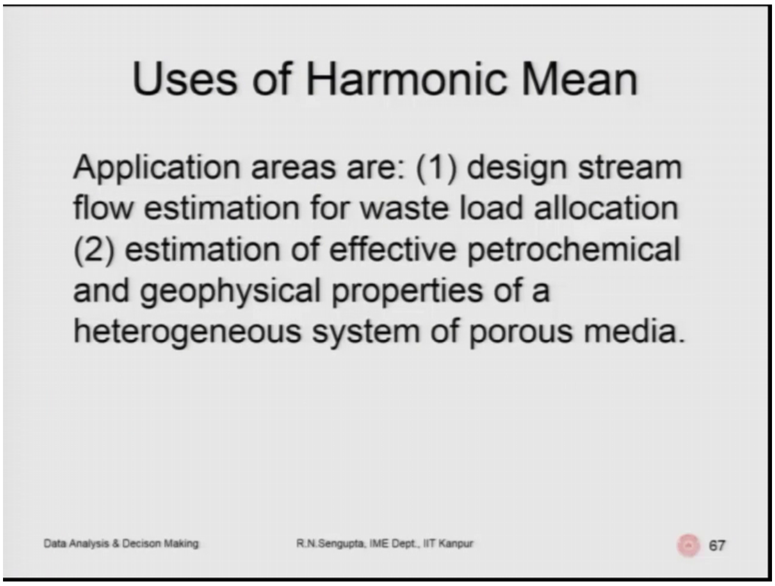
So, consider car travels at distance x with velocity v 1. Consider I am going from Delhi to Jaipur or I am traveling from Bangalore to Bengaluru to Chennai. Consider car travels a distance x with the velocity v 1 and returns back; that means, it returns back from Chennai to Bengaluru .

The same distance with the velocity v 2, so we and car and we want to find out what is the average velocity so; obviously, the in the first case the time taken is x by v 1 which is distance by velocity and the next instant when it returns back the time taken is x by v 2.

So, the total time taken is x by v 1 plus x by v 2 which is in the denominator as you can see in the slide. Now I want to find the average velocity. So, average velocity would be basically with the total distance traveled divided by the total time, total distance travel is that means, I have traveled from Bengaluru to Chennai and from Chennai back to Bengaluru so, total distance is 2 x and the time taken as i mentioned is x by v 1 plus x by v 2.

So, we can find out the corresponding velocity. So, in this case you will find out that that we are using trying to as. So, if you want to find out the average velocity, so x both from the numerator and denominator will cancel. So, it will basically be 1 by x v 1 plus 1 by v 2 divided by 2 so; that means, this is the way how you calculate the harmonic mean. Hence we see that in this case we use the harmonic mean and not the arithmetic mean to find out the average of the speed that is just an example.

(Refer Slide Time: 02:53)



## Uses of Harmonic Mean

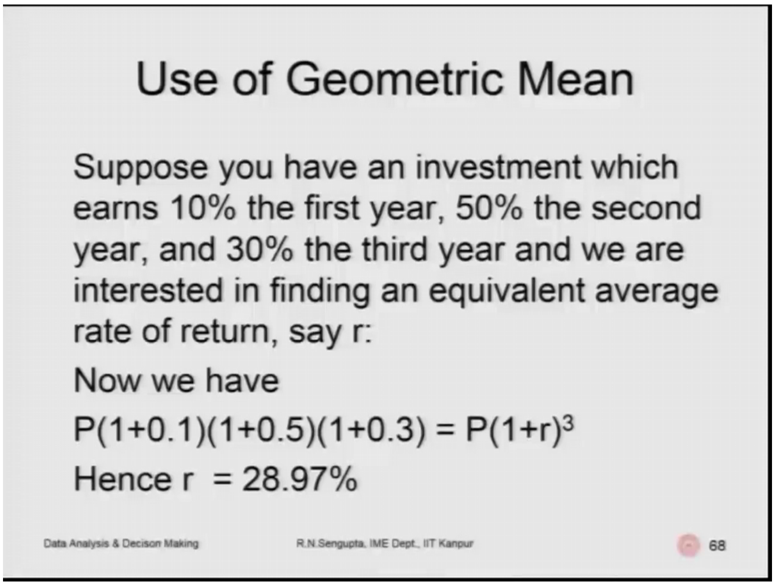
Application areas are: (1) design stream flow estimation for waste load allocation (2) estimation of effective petrochemical and geophysical properties of a heterogeneous system of porous media.

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 67

So, use of Harmonic mean they can be applied in areas like design of stream flow estimation for waste load allocation, number 2 estimation of the effective petrochemical and geophysical properties of a heterogeneous system of porous media and also the example which we considered for finding on the average speed between 2 two cities or it can be 3 cities, 4 cities also depending on how the problem have been formulated. So, in case it is 3 cities and consider for the time being the distances are the same.

So, you are traveling from city 1 to city 2 distance x, city 2 to city 3 distance x, city 3 to city 1 distant x so; obviously, it with 3 x divided by in the first case it will be x by v 1 then x by v 2 and x by v 3 three. So, xx cancels and the value would be 3 divided by in in numerator is 3 and then denominator it will be 1 by v 1 plus 1 by v 2 plus 1 by v 3. Now we will consider the case of geometric mean, I have already given the example I will just repeat it for your own convenience.

(Refer Slide Time: 03:59)



Use of Geometric Mean

Suppose you have an investment which earns 10% the first year, 50% the second year, and 30% the third year and we are interested in finding an equivalent average rate of return, say r:

Now we have

$P(1+0.1)(1+0.5)(1+0.3) = P(1+r)^3$

Hence r = 28.97%

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 68

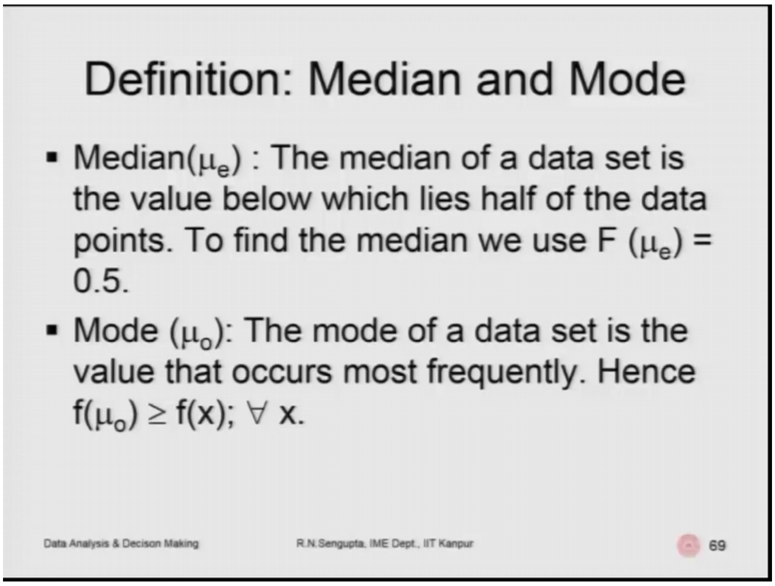
Suppose you have an investment which earns 10 percent in the first year, 50 percent in the second year, and 30 percent in third year and we are interested to in finding the equivalent average rate of return consider is r for all the 3 years. So, in the first year if the principal amount is P, the total amount of value of money which you have after the interest rate is added it to be P multiplied by 1 plus 0.1 in the, after two years it will be principal amount into the interest rate which you earn which is 1 plus 0.1 which you already know.

In the second year, it is 1 plus 0.5 which is for the second year only for the second year. Now if you have basically want to find out for the third year, the interest rate would basically be 1 plus 0.3. So, the overall total value would be P into 1.0 plus 1 plus 0.2 plus 1 into 1 plus 0.3 sorry not plus multiplied and if I if I think that equivalent interested is r.

So, it will be P into 1 plus r to the power q. So, in this case we can find out the overall interest rate based on which you can do the calculation comes out to be about 28.97, which is about 29 percent, so this would be the average interest rate which we calculate using the concept of geometric mean because why geometric mean because what do you are doing is that you multiplying the interest rate 1 plus 0.1 into 1 plus 0.5 into 1 plus 0.3 multiply all of them and find out the one-third cube the cube root of this.

So, it will be one-third of this of this value to the power one-third means not in the in the division sense, so that will that value comes out to be as I mentioned 28.97.

(Refer Slide Time: 05:42)



## Definition: Median and Mode

- Median($\mu_e$) : The median of a data set is the value below which lies half of the data points. To find the median we use F ($\mu_e$) = 0.5.
- Mode ($\mu_o$): The mode of a data set is the value that occurs most frequently. Hence $f(\mu_o) \geq f(x); \forall x$.

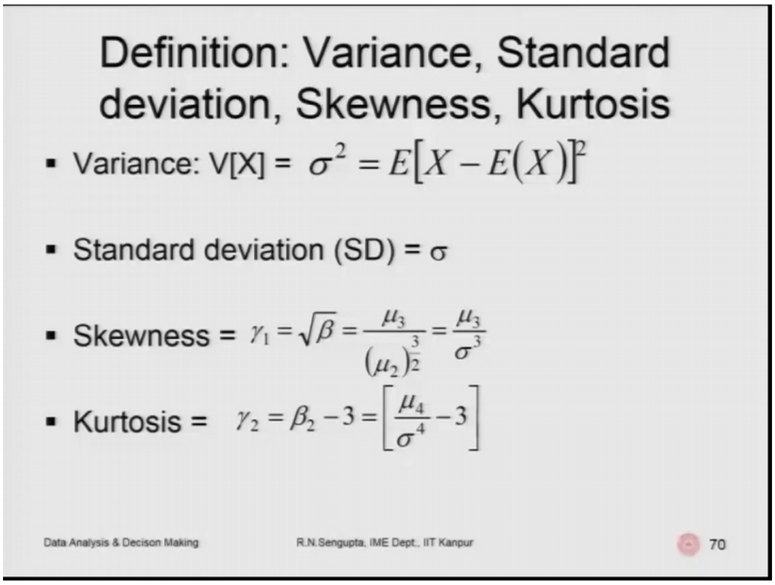
Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 69

Now, as I have been mentioning about the median and the mode even though I did not mention about the mode many times that I did mention about the median. So, I will just try to give the definitions of the median and the mode. The median which is denoted by mu suffix e, because if you remember when we discussed the mean value, it was only given by the in the mu without any suffix. So, it if the median would be denote by mu suffix e, the median of our data set is the value below which lies half of the data point and above which will has the half of the data point.

That means it trying to seen divided the overall data sets corresponding in the probability 2 equal halves. So, that is why the (Refer Time: 06:17) which we found out the midpoint divides the total frequency set of values of frequencies, not the values in frequency into 2 equal halves. To find out the mean, we would basically pour find out the overall sum of the probabilities from the minimum value to that value has the probability is 0.5.

So, that is that is why you equate it to 0.5 while mode is the value of a data set which is with is with that occurs the most frequently the maximum number of times hence the mode of the would basically we found out for those values for which the frequency or the relative frequency on the probability is the highest. So, small f of mu naught, so now, here the mode is denoted by mu symbol with the suffix naught to basically denote the mode. So, small f suffix mu naught should always be greater than equal to whatever the x values are there on to the left or the rights, so the maximum value.

(Refer Slide Time: 07:13)



## Definition: Variance, Standard deviation, Skewness, Kurtosis

- Variance: V[X] = $\sigma^2 = E[X - E(X)]^2$
- Standard deviation (SD) = $\sigma$
- Skewness = $\gamma_1 = \sqrt{\beta} = \frac{\mu_3}{(\mu_2)^{\frac{3}{2}}} = \frac{\mu_3}{\sigma^3}$
- Kurtosis = $\gamma_2 = \beta_2 - 3 = \left[\frac{\mu_4}{\sigma^4} - 3\right]$

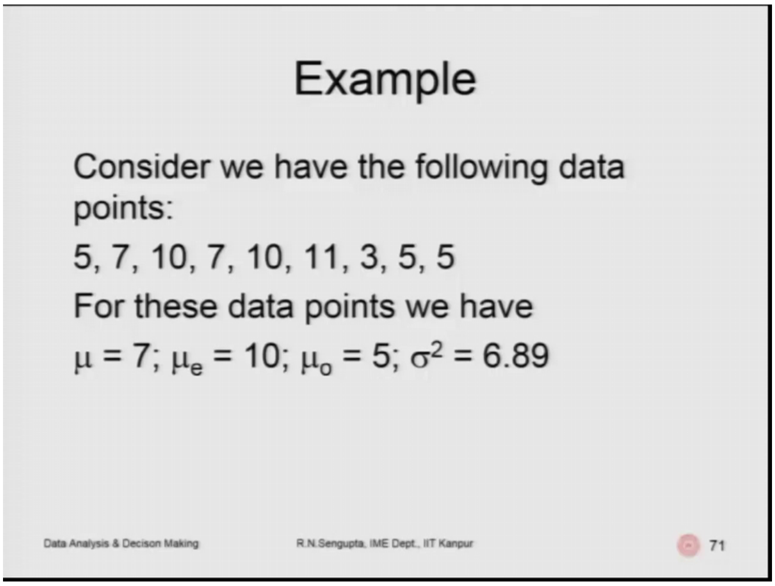
Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 70

. So, now, definitions of the variance which is the some dispersion concept, the variance standard deviation Skewness and Kurtosis the variance values given by this the symbol sigma square and that is calculated by the expected value of the differences between the actual value X minus is expected value. So, what we do is that you find out the difference between the and so called value with its average value square them up and basically find out the expected value for that.

So, that will give me the variance and if I talking about the standard deviation it will be the square root of variance, while Skewness and Kurtosis have their corresponding values are given in this equation. So, now, remember here this mu suffix 3, mu suffix 4, mu suffix 2 are basically the mean value corresponding to the third moment. So, these are the mean variance for the third moment not corresponding to the third moment by itself these are the mean values corresponding with the third moment and we will find out the values and put it replace them in the Skewness and Kurtosis and do the calculations accordingly.

So, let us consider a very simple example we have the following data points.

(Refer Slide Time: 08:25)



## Example

Consider we have the following data points:

5, 7, 10, 7, 10, 11, 3, 5, 5

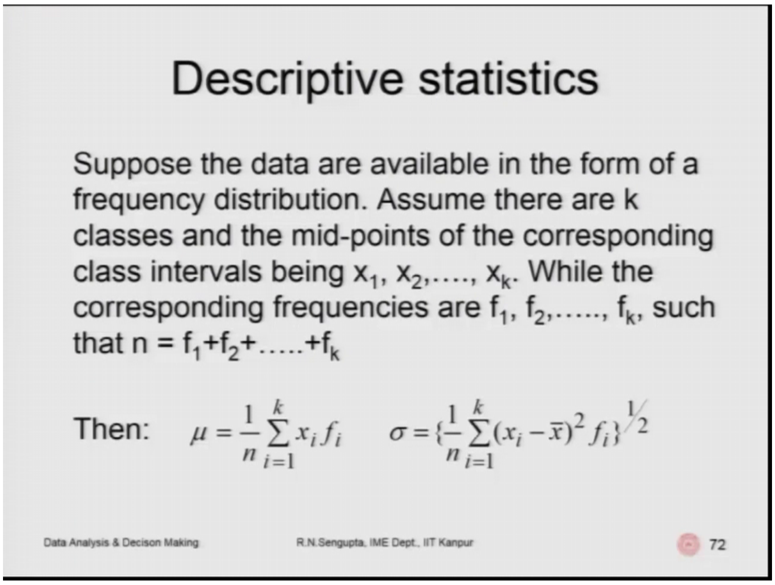
For these data points we have

$\mu = 7$; $\mu_e = 10$; $\mu_o = 5$; $\sigma^2 = 6.89$

Data Analysis & Decison Making    R.N.Sengupta, IME Dept., IIT Kanpur    71

Which are 5, 7, 10, 7, 10, 11, 3, 5, 5. So, they are not arranged according to starting from the minimum to the maximum or maximum to the minimum, for these points we can find out the mean value. So, what how we find out the mean value is basically add them up divide by the total number of readings. So, the total number of readings is 9. If I want to find out the median, you will find out the median, so as that it basically divides the that dataset into 2 equal halves the mode would basically with the highest value. So, in this case we find out number 5 occurs the maximum number of time is 5 and the sigma square values would be calculated corresponding to the formula, which is the expected value in the bracket x minus expected value x whole square and basically we find out the values comes out to be 6.89.

(Refer Slide Time: 09:10)



## Descriptive statistics

Suppose the data are available in the form of a frequency distribution. Assume there are k classes and the mid-points of the corresponding class intervals being $x_1, x_2, \ldots, x_k$. While the corresponding frequencies are $f_1, f_2, \ldots, f_k$, such that $n = f_1+f_2+\ldots+f_k$

Then: $\mu = \frac{1}{n}\sum_{i=1}^{k} x_i f_i \qquad \sigma = \{\frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i\}^{1/2}$

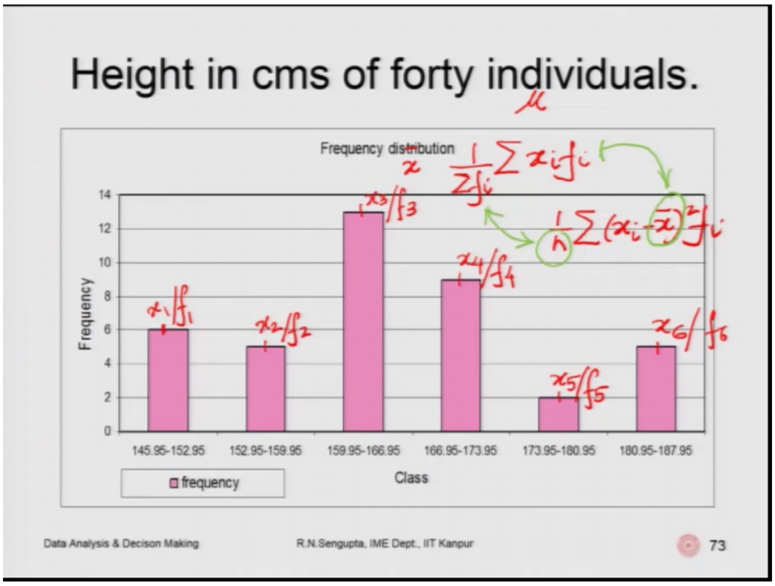
Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 72

Suppose the data are available in the form of frequency distribution assume there are k classes and the midpoints of the corresponding class intervals are given as x 1, x 2, till x k so; obviously, we have class interval. So, the intervals have in such a way that the midpoints are given. So, while the corresponding frequencies of those classes I given as f 1 to f k then the total frequency is basically sum of all the values, which is f 1 plus f 2 plus f 3 till the last value which is f k. If I want to find out the overall mean so; obviously, it would mean than I am basically multiplying the mid value which basically is the characteristics of the interval multiplied these corresponding frequency. So, it will be x i into f i sum them up and divide by n, which is basically the total frequency.

So, what I am trying to do is that find out the multiplication, multiplication value of x 1 into f 1 plus x 2 into f 2 plus x 3 in to f 3 till the last value which is x k into f k sum them up and divide by n, n is basically the sum of f 1, f 2, f 3, till f k. So, that will give me the mean value. If I want to find out the variance for that it would be again I am trying to utilize the same formula it will be x i which is the middle value which I have considered x 1, x 2, x 3 till x k minus the mean value which you have just found out.

So, now you will see there is a difference between mu and x bar is basically I am trying to denote that is for one is for the population and one is for the sample and come to that definition later on, technically just remember mu is for the population and x bar is basically for the sample. So, coming back to the formula which is x i minus x bar whole square multiplied by the corresponding frequencies because frequencies have to we will utilized in order to find out the variances then find and obviously, you will you will as it is standard deviation hence the 1 by 2 power the square root is coming, but technically this squared. So, x i minus x bar whole square into f i sum them up divided by n which is the sum of all the probabilities. The frequencies from f 1 to f f k and find out the square root that will be give me the standard deviation.

(Refer Slide Time: 11:17)



Height in cms of forty individuals.

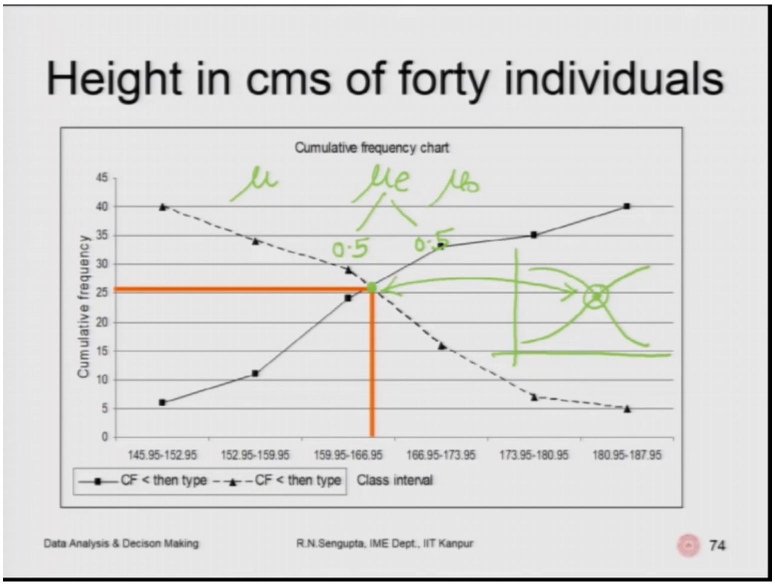
Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 73

So, coming back to the example of the height. So, now, the frequencies with the class intervals are given 145.95 to 152.95, the second one being 152.95 to 159.95, the last one is 182.95 to 180.95.

So; obviously, if I want to find out the overall average value what I will do is like this, I have the midpoint values here. So, this is x 1, this is x 2, this is x 3, this is x 4, this is x 5, this is x 6. So, the frequencies which I find out is 6 where I am hovering my pen, I am not writing it down. So, let us let it be f 1, second value is about 5 So, this would be f 2, third value is 13. So, it will be f 3. I am not writing the values i am just writing this the variables of the symbol. The fourth frequency is basically 9, which is f 4,fifth value is 2, which is f 5, sixth value is 6, which is f 6. So, what I do is that sum the values x i into f i divided by the summation of f i and that would give me the mean value which I will denote by x bar for the sample, if it is not for the sample it will denote by mu, but I am not going to consider that then the variance would be given by the value of 1 by n summation of x i minus x bar whole square into f i.

Now, this one this value of x i which i find out is basically coming from this equation. So, and this one f n is equal to this one which I will considered. So, based on that I find out the stem the variance square root of that will give me the standard deviation.

(Refer Slide Time: 13:42)



Height in cms of forty individuals

Data Analysis & Decison Making

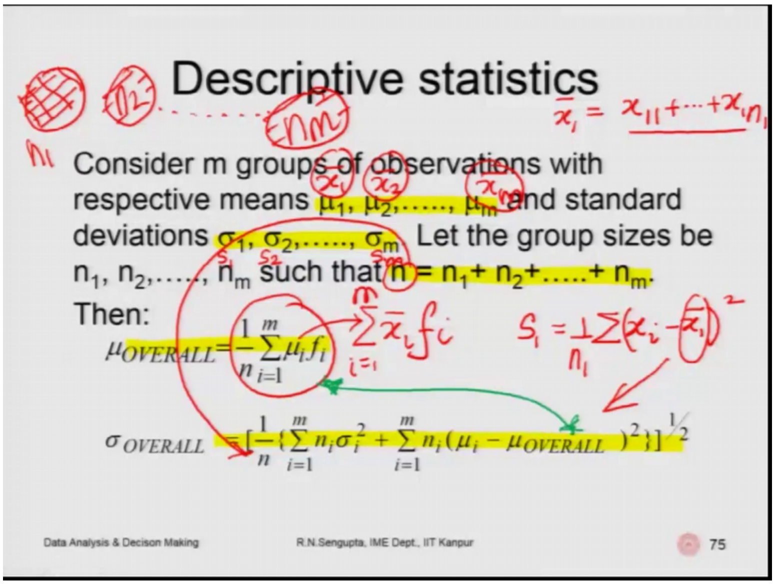
R.N.Sengupta, IME Dept., IIT Kanpur

74

. So, again the cumulative frequencies are given. So, this is the mean the median value which means if I am able to draw those j curves.

So, this is the value based on which we I am doing the calculations. So, the mean of the median value is given by mu suffix e, the mean is given by mu, mode is given by mu naught. So, this mu suffix e value we did divide the overall probability in distribution such a way. So, the probability is go to the left on the right are 0.5 05 point five.

(Refer Slide Time: 14:26)



n1
n2
nm
$\bar{x}_1 = x_{11} + \cdots + x_{1n_1}$
$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_m$
$s_1$
$s_2$
$s_m$
$\sum_{i=1}^{m} \bar{x}_i f_i$
$s_1 = \frac{1}{n_1}\sum(x_i - \bar{x}_1)^2$

# Descriptive statistics

Consider m groups of observations with respective means $\mu_1, \mu_2, \ldots, \mu_m$ and standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_m$. Let the group sizes be $n_1, n_2, \ldots, n_m$ such that $n = n_1 + n_2 + \ldots + n_m$.

Then:

$$\mu_{OVERALL} = \frac{1}{n}\sum_{i=1}^{m}\mu_i f_i$$

$$\sigma_{OVERALL} = [\frac{1}{n}\{\sum_{i=1}^{m} n_i \sigma_i^2 + \sum_{i=1}^{m} n_i(\mu_i - \mu_{OVERALL})^2\}]^{1/2}$$

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 75

Now let us consider m groups observations with respect to means of mu 1, mu 5 till mu m.

So, this mu 1 to mu m I am considering they are from the population, but if they are interchange with the sample observations they would be corresponding to let me change the color this is x 1 one bar, x 2 bar, x m bar and the standard deviations are given technically these values and use a highlighter these values which are there are for the population like these are for the population, these are also for the population.

So, now the color scheme which I am using is the yellow one is for the population and the red color which I am using is for the sample, sample, sample corresponding values I will denote by s 1 for the standard deviation, which is basically word is used standard deviate. I am going to come to that later on for the first set s 2 for the second set till the last one which is s m so; obviously, x 1 bar would be the mean value for the first group.

So, it technically it will be x 1 1, the first 1 is for the first set, the second 1 is for the observations. So, consider there are k number of observations. So, this k may change from group to group and s 1 would be calculated considering I find out and I do my calculation it recorded. So, let the group size be n 1 to n n m. So, this is what k was i was talking let me change it. so for no confusion should be there. So, this should be n 1.

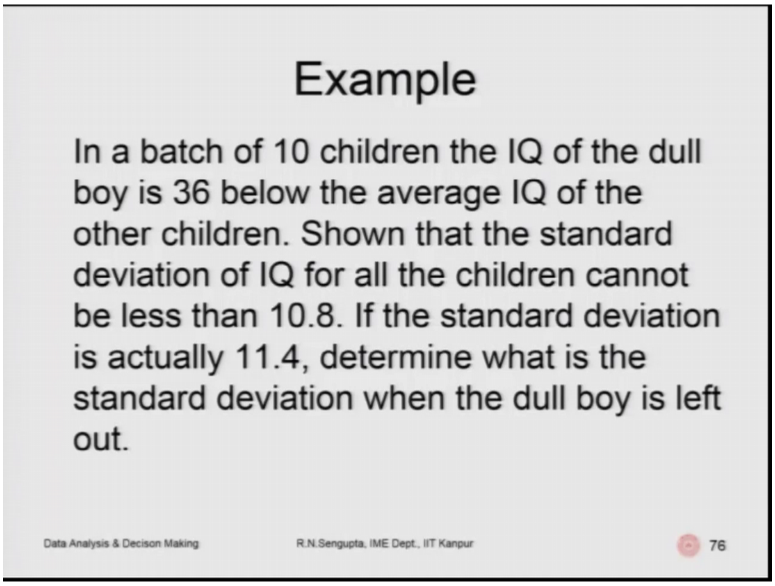
So, n 1 so, this observations are first group has n 1, second group has n 2 observations, the last one has n m observation. So, there m naught of course, and the sum from x 1 for the first one to the mth 1 are given by sum of n 1, n 2, n 3 till last one. So, this is what try to again highlight using the yellow color. So, if you see the over overall mean value. So, this is the mean value which is mu i which would be x 1 bar depending on whether using the sample of the of the population multiplied by the frequency sum them up for all the values of x equal to 1 to m. So, this m is basically for this am I again denoting.

So, if i want to find out the overall variance, so over all variance would basically be considering the individual variance and the group wise varies and the overall one.

So, it will be 1 by n where n is the total set which is here. So, this is coming here. So, remember that multiplied by in the bracket n i into n i is n 1, n 2, n 3 till n m multiplied sigma i square, i square for the corresponding groups plus the difference between the average value if you remember. So, this one so, this is being replaced by n i into mu i in the bracket mu i minus with the mu overall.

So, mu overall is 1 and change the color scheme or else it will become difficult. So, this is going to come here. So, based on that i square them up find out the overall value. So, this gives me the overall standard deviation, this gives me the overall value, average value.

(Refer Slide Time: 19:44)



## Example

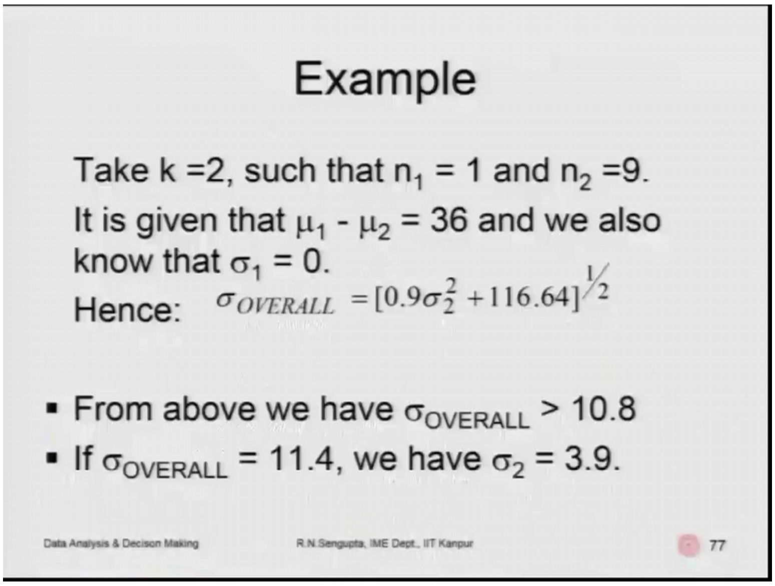
In a batch of 10 children the IQ of the dull boy is 36 below the average IQ of the other children. Shown that the standard deviation of IQ for all the children cannot be less than 10.8. If the standard deviation is actually 11.4, determine what is the standard deviation when the dull boy is left out.

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 76

In a batch of 10 children the IQ of the dull student is 36 below the average IQ of the other children. So, this is just a simple problem which you are trying to utilize using the concept which you have just studied. We are required to show that the standard deviation IQ for all the children cannot be less than 10.8.

So, if the standard deviation is actually having 11.4 determine what is the standard deviation when the dull boy is left out. So, now, we have 10 children and the IQ of the dull children is 36 below the average IQ of that of the other children. So, let us divide two groups, which is m is equal to two in the first group or in the second group wherever you put, there would be 9 students and 1 students which is the dull one based on that we will do the calculation. So, let us proceed.

(Refer Slide Time: 20:38)



## Example

Take k =2, such that $n_1$ = 1 and $n_2$ =9.
It is given that $\mu_1 - \mu_2 = 36$ and we also know that $\sigma_1 = 0$.
Hence: $\sigma_{OVERALL} = [0.9\sigma_2^2 + 116.64]^{1/2}$

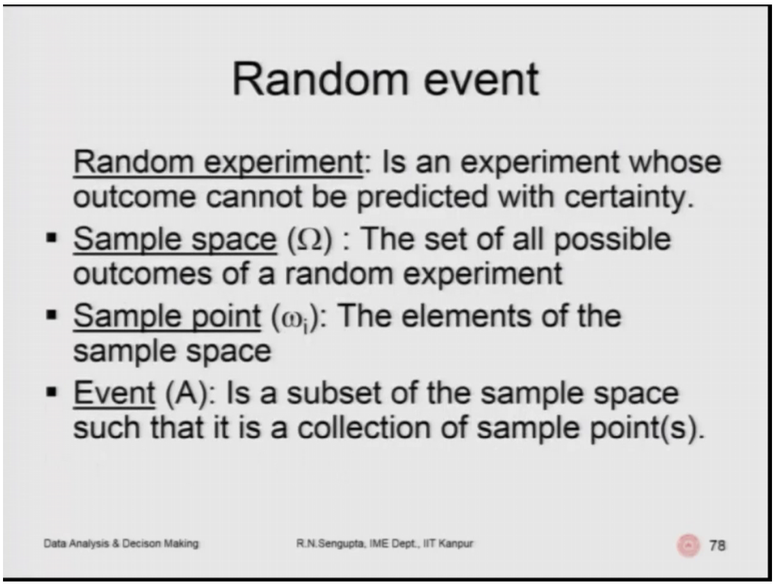
- From above we have $\sigma_{OVERALL} > 10.8$
- If $\sigma_{OVERALL} = 11.4$, we have $\sigma_2 = 3.9$.

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 77

Let us k take k has 2 such that n 1 which is the dull student in the observation size is 1, for the other group observation size is 9 because there are total ten students, it is given the difference between the mean value for the IQs between the dull student and the rest is 36. So, it will be mu 1 minus mu 2 is 36 so; obviously, it is the dull students would be lower and we also known the standard deviation for the only 1 student being there is 0 standard deviation for the other group; obviously, would not be 0 because this is 9.

Based on that if we do the calculations we find out the overall comes out to be 0.9 sigma 2 squared which is for the second group where there are 9 students plus 116.64 whole square root of that from the above we have the overall standard deviation has to be greater than 10.8 as we proved and if the overall value is 11.4, we find out the standard deviation for the group which is 9 student comes out to be 3.9.

(Refer Slide Time: 21:44)



Random event

Random experiment: Is an experiment whose outcome cannot be predicted with certainty.

- Sample space ($\Omega$) : The set of all possible outcomes of a random experiment
- Sample point ($\omega_i$): The elements of the sample space
- Event (A): Is a subset of the sample space such that it is a collection of sample point(s).

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 78

Now I will discuss something about the random event and few a few definitions of that for. So, let us go it slowly and try to understand that. So, what we mean by random event. So, random experiment so, in an experiment with the outcome cannot be predicted with certainty like tossing the coin rolling the die or what is the probability that a car would be going through this gate considering there is some distribution for the car flow or consider that they would be rainfall consuming considering there we have the information set for the distribution as such.

Or the telephone calls are doing or how much time the tailor machine will take in trying to process each and every person who is approaching the tailor if the machine the machine is the dispensing machine or the cash or the counter from where you are trying to basically either depositor check in cash or check update your passbook and whatever it is. So, in this experiment the outcome cannot be predicted with certainty, but you would definitely have some underlining distribution to do that analysis we will come to that later on.

Now the samples space which is the capital phi is basically the set of all possible outcomes of the random experiment which is fall into that for that experiment, now let me give you an example consider you are tossing a coin and we are, we want to toss the coin only twice not more than that. So, in that case the overall sample space can be as follows. If the first toss is tail consider the second toss is also tail.

So, it is tail tail, so that is one of the element for the sample space first the first toss tail, second toss head, so it is tail head, third example is head tail and the fourth example is head head. So, if we have this four points basically makes the sample space based on which you can do the experiment. Now consider a different above example I say and we will consider these examples later on I say that let us consider the sample space till we get the first head.

So, the expect the sample you put like this the points or the sample space now technically is not finite, but infinite why because if the head comes in the first toss that is one of the point what if the head does not come in the first toss it will be tail head, what if the head does not come in the first two tosses, it will be tail tail head, what if the head does not come in the first three tosses, it would be tail tail tail head if you basically think, it can basically be extended to infinity such that the overall sample space now would be a head, tail head,, tail tail head, tail tail tail head and basically going to infinity. So, we will basically consider the case where the sample space is infinite. The sample points would be the elements of the sample space. So, in the example which we considered for the sample space which is finite which was basically head tail tail tail tail head head head.

So; obviously, these four would be the distinct sample points which we have and if you considered it very simply and think then the corresponding probabilities in all the 4 cases when there is a finite sample space would be half into half, it will be one-fourth. considering all the four points which are there; that means, tail tail tail head head tail head head all have the equal probabilities because tossing a the coin the probabilities of getting a head or a tail are equal that is half and half. So, hence it is half into half for all the cases one-fourth.

Now, consider the next example which I just gave few seconds back. If you want to basically stop till the time you get the first head now the observed sample points are like this head, tail head, tail tail head, tail tail tail head and so on and so forth. So, in that case the corresponding probabilities are not equal because for the head probability is half, for the probability of tail head is half into half, probability of a tail tail tail for a sorry tail tail head is half into half into half, probability of a tail tail tail head is half into half into half into half.

So, you will see the corresponding sample points are also infinite and the corresponding probabilities are also infinite. Now consider we want to find out what is an experiment and we will define an event, depending on the experiment it will be the subset of all the samples space which you have set of all those sample points which you have such that it will basically define the exact event which you have.

So, say for example, I want to find out the probabilities in the first example that I get at least 2 tails, so in that case if it is at least 2 tails; obviously, the event would be tail tail. So, any event which is head tail tail head or head head is not counted. If I am saying that the or here let I have missed something. So, if it is at least 2 heads; obviously, it would be in that case would be a head would have also come like a tail head or a head tail so; obviously, there would be three such conglomeration of sample points which will give me the event.

Now in the other example which I said that if I find want to find out that the head comes in the second at least in the second throw the first and the second would be counted and if I consider the example that we have to basically tossed it 5 times or 10 times depending on what is the outcomes; obviously, the sample points would change sample points for that event would change because the sample points to the overall sample space remains the same such that we will define the event accordingly.

So, with this I will close this fourth lecture and continue the discussions in the fifth lecture about sample points examples and how we can basically bring that concept in the example of probability and discuss it further. Have a nice day and.

Thank you very much.