

Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 03
Arithmetic mean

Hello and welcome to today's class. In our last lecture I had discussed about the type of data you acquire from a given experiment and the ways and means of presenting that data. So, in today's class, we will see how you from that data you can extract some quantitative parameters to distinguish between 2 experimental conditions and say whether those differences are significant or what kind of conclusions can you draw by looking at the data.

So, before we get started I would like to do a brief recap of what we had discussed in last class.

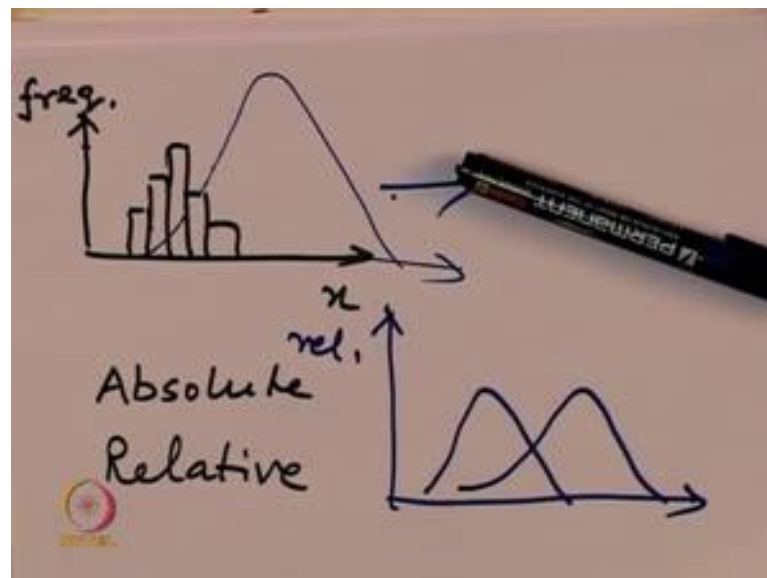
(Refer Slide Time: 01:07)



In the last class I had discussed about the types of data, which would mean categorical data which is more; you know qualitative in nature or quantitative data even within the types of data in quantitative data you might have discrete variables like the number of visits to a clinic so on and so forth or continuous variables like the time it takes to for a student to solve a problem.

Then we discussed about the types of plotting and so, expressing for you know standard representation of qualitative data is used is using pie charts or bar charts, but the caveat is that these kind of charts you can only plot when the number of conditions or the number of you know distribution is not widely distributed in other words if you having a pie chart if you have to show 200 conditions then it is impossible to fathom any sense from that plot because everything has to be represented in terms of a circle then we talked about the histogram it is one of the most widest used distributions and how you go by converting data raw data by using frequency tables you know you make bins you make frequency tables. So, that you can put that data into some form and you then you plot in terms of values and their frequency and you get what is known as a histogram.

(Refer Slide Time: 02:23)



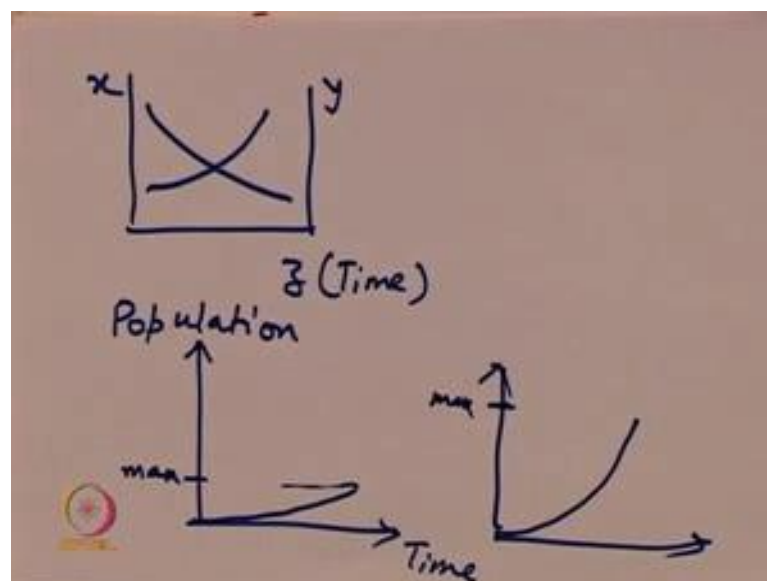
So, this is a sample representation of a histogram right we can have these are this is my frequency axis this is my variable x , let us say this vertical axis my frequency and this is how I can plot them now these frequency axis can be plotted in absolute terms or in relative terms and this relative frequency is important when you are comparing between 2 separate experiments or 2 different conditions and the number of data points is not the same let us say in one case you have obtained the data for 50 total of 50 points and the other case there are 100 points.

So, if you were to put that in the same frequency data the plot might look. So, if I were to overlay on this existing plot a different plot it might look like this, but in reality if you

normalize. So, if you actually normal, if you normalize this data so that what you might have your data might actually look like this. So, this you have relative frequency where you divide with respect to the maximum the total number of observations.

After histogram, we discussed about line and scatter plots these are widely used. So, you just you know depending on your variables let us say you have 2 variables x and y and you are plotting how y is varying with respect to x you can put them as scatter plots where they are individually point you know each point is the x y representation in 2 d, but of course, with scatter points you can convert it into line plot only if it makes sense if these data have some very clear trend that it can be approximated by a line then it makes sense to make it as a line plot I also briefly discussed about this double y plots where you have to. Let us say you have 3 variables x y and which are variable with which are changing with respect to a variable z then you can plot.

(Refer Slide Time: 04:26)



Then you can plot; let us say x and y with respect to your variable z and let us just say that as I ; let us take a simple example where z is time and x and y are 2 variables which are varying with time and we want to see how x and y correlate.

So, as per the plot I have drawn as x decreases. So, in our case, there is my time axis at x decreases with time y increases with time which tends to say that they are inversely correlated. Next thing I wanted to talk about, in terms when you are making the plot few

things to keep in mind you must label your axis appropriately you must choose appropriate units.

(Refer Slide Time: 05:04)

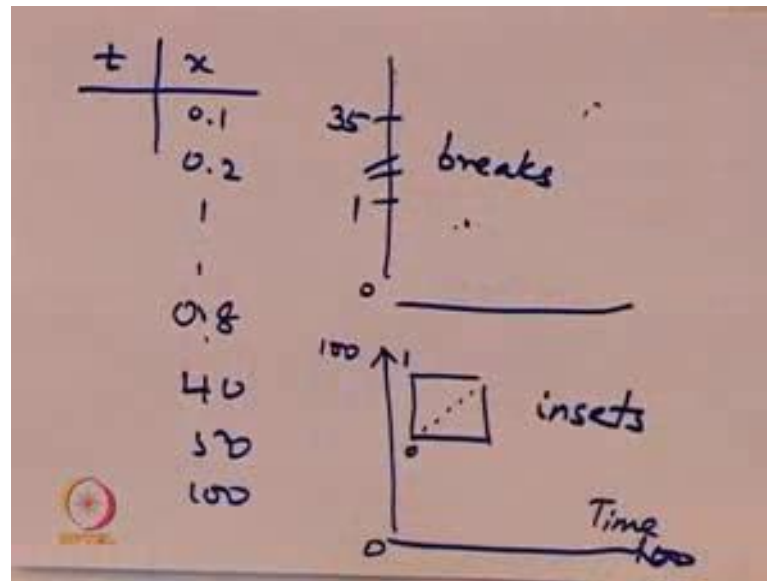


For example, let us say we are talking about cell speed you can report or you know any speed you can report in micron per minute micron per hour meter per minute meter per hour so on and so forth depending on what you are measuring and what is the minimum amount you can measure in accurate manner.

So, that also brings us that you want to choose the appropriate range let us say you are seeing a population explosion and you the maximum value you choose for your y axis. So, if I go to the example let us say this is time and this is my population if I want to convey that my population is really increasing with time and I choose my y range such that my curve only looks like this then what I want to convey does not flow with the way I have plotted the data much rather what I should do is I should plot the same curve with.

So, if this is my max value I can choose this max value as reasonably high. So, I can really see a non-linear growth in the way the population is expanding as opposed to choosing an arbitrary large value for the y axis.

(Refer Slide Time: 06:27)



The other thing important thing to note is when you have large variations in data let us say most of your values I am just plotting this is my time axis and this is my x axis and x values are you know either you have very low values point 1, point 2, 1 and then suddenly the next value goes to you know 40, 50, 100. So, it is very difficult to represent all these values of x in the same plot. So, there are 2 approaches. So, what you do you either put an insert a break in the axis. So, let us say my 0 to 1 is in one graph, I and the next point starts is from 35. So, my 40 point will raise somewhere here and my you know let us say if this large values 0.8 it will be somewhere here. So, I can see both these points on the same axis.


The alternative to that is to choose is to plot over a wider range and what you do is you show a smaller range let us say this is from 0 to 100 and this is from 0 to sorry, this vertical axis is from 0 to this; this is your time axis. So, this vertical axis is from 0 to 100, but the inset is only from 0 to 1. So, here you can have these extra points and then you can clearly see how they vary in y axis. So, this is using insets and this is using breaks.

(Refer Slide Time: 07:48)

Example 1

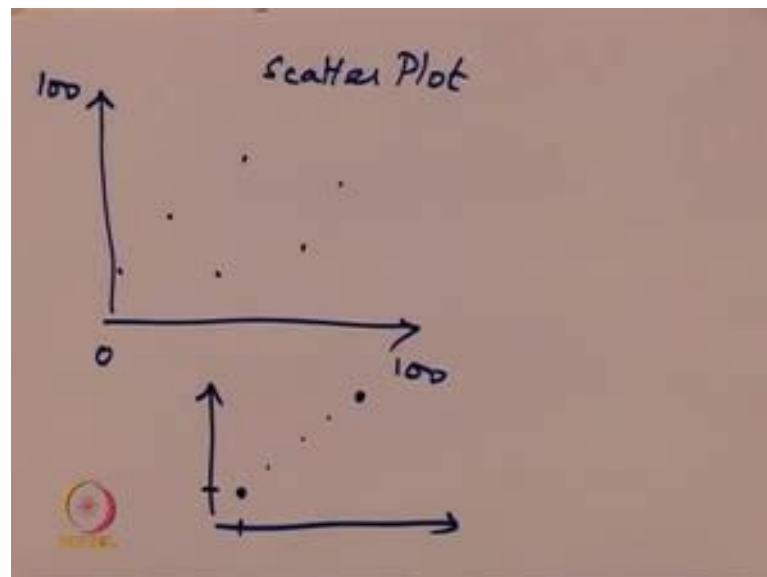
Test Scores of 10 students in 2 exams

61	49
93	74
87	66
42	45
55	68
67	88
82	59
50	71
29	21
55	50



So, let us just begin today's lecture with one example. So, this is a sample representation of data of 10 students in 2 exams and we want to find out how we should represent this data. So, I presume it is reasonably clear to most of you that since it is x and y for 10 students then what you can do is the easiest way to plot this data is using a scatter plot.

(Refer Slide Time: 08:09)



So, you can plot using scatter and for each point you can see is a 61 and 49 and so on and so forth. So, let us say your maximum value is in the range of 0 to 100; 0 to 100 and then accordingly you can put individual points and see how these performances are correlated.

So, your aim might have been to see that how the students are performing across different exams and these are in 2 different exams to perhaps 2 different subjects. So, if you see a very positive correlation.

So, let us just say one simple case that you have sorted this data and you have plotted them separately such that. So, here I have just randomly plotted, but you have plotted in such a way that the student. So, this is the student with the lowest score in exam 1, he or she also scores the lowest score in exam in exam 2. So, they have the lowest score. So, if a curve if you would sort the data and re plot this it conveys the message that the students have reasonably they vary in their standards and the student who is good is consistently good and that is the reason why high score in exam 1 also translates into a high score in exam 2. And, similarly for the student who is weak low score in exam 1 is also closely related to a low score in exam 2.

(Refer Slide Time: 09:42)

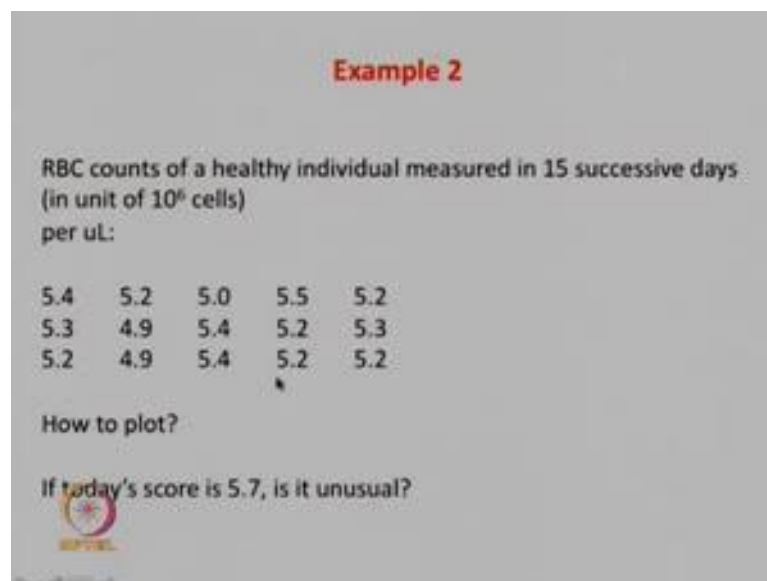
Example 2

RBC counts of a healthy individual measured in 15 successive days
(in unit of 10^6 cells)
per μL :

5.4	5.2	5.0	5.5	5.2
5.3	4.9	5.4	5.2	5.3
5.2	4.9	5.4	5.2	5.2

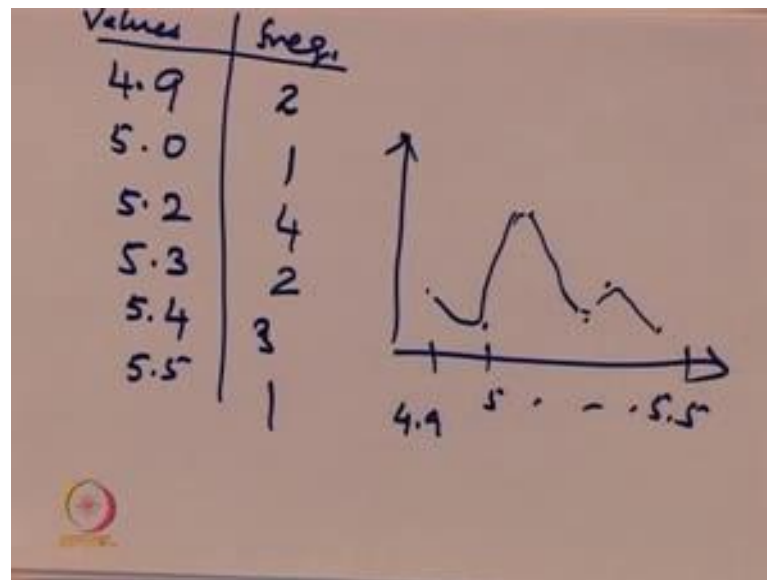
How to plot?

If today's score is 5.7, is it unusual?



So, let us take another example. So, we now have this is the data for RBC counts of a healthy individual measured in 15 successive days. So, is it has units of 10 to the power cells per m l per micro litre u litre is short form for micro litre. So, what you can see is you can have these important values. So, the question is how should we represent this data?

(Refer Slide Time: 10:07)



So, if I were to look at the values which are occurring I have 4.9, I have 5, I have 5.2, I have 5.3, I have 5.4 and I have 5.5. So, these are the values and I can plot their frequency 4.9 appears twice, 1 appears once, 5.2 appears 1, 2, 3, 4, times, 5.3 appears 2 times, 5.4 appears 3 times and 5.5 appears once.

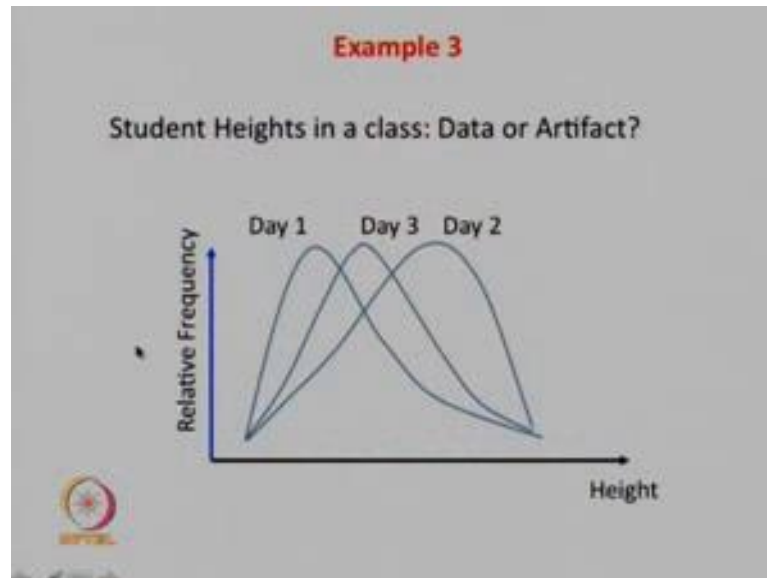
So, the very fact that I have actually converted into frequency plot would mean that eventually you would plot this data as a histogram. So, these are your individual values 4.95 so on and so forth. This is a 5.5 value and what I can see is, if I were to, this is my 2 1, then I have a maximum 4, then I have 2 3 1. So, if I were to actually connect these data. So, my curve would look something like this. So, the histogram looks something like this.

Now, whether or not it is; this is too few numbers. So, it is difficult to comment on it, but if let us say I asked you then following question let us say in today's score. So, this is the; your values for the last several days and today's score is 5.7 is it unusual. So, you can clearly see from your ex you know existing histogram that there is no 5.7 value. So, one would argue that 5.7 is a little bit unnatural maybe you are you know suffering from some disease or infection. So, it is it warrants a doctor's examination.

So, let me take another example. So, imagine we have you know I actually take the statistics of average student's height in a class and I do it on 3 successive days and this

is. So, this is the plot of height and relative frequency on day one this is my distribution on day 2 this is my distribution and on day 3 this is my distribution.

(Refer Slide Time: 12:24)

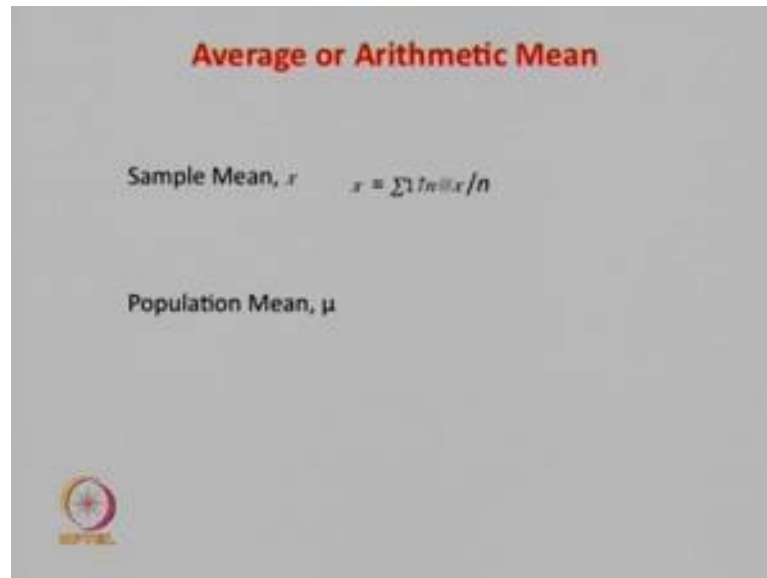


So, this is you know mind you it is not from 3 different populations it is from the same population right. So, when the; you know it is a little bit baffling as you can see that there are some overlapping regions, but the curves look reasonably different from each other. So, what are the qualitative aspects of these curves which are different what I can clearly see is if I were to look at the maximum in peak height. So, for which you have the maximum number of observations it is at this value to the left on day 1 to the right on day 2 and to the middle on day 3.

Now, if I were to imagine that on an average your population consists of boys and girls and on an average boys are you know are taller than girls then it is possible that on day 1, you had a greater number of girls present in the lecture on day 2, you had a greater prevalence of boys attending the lecture and on day 3, there is a reasonably equal distribution or equal fraction of the boys or the girls who were present in the class.

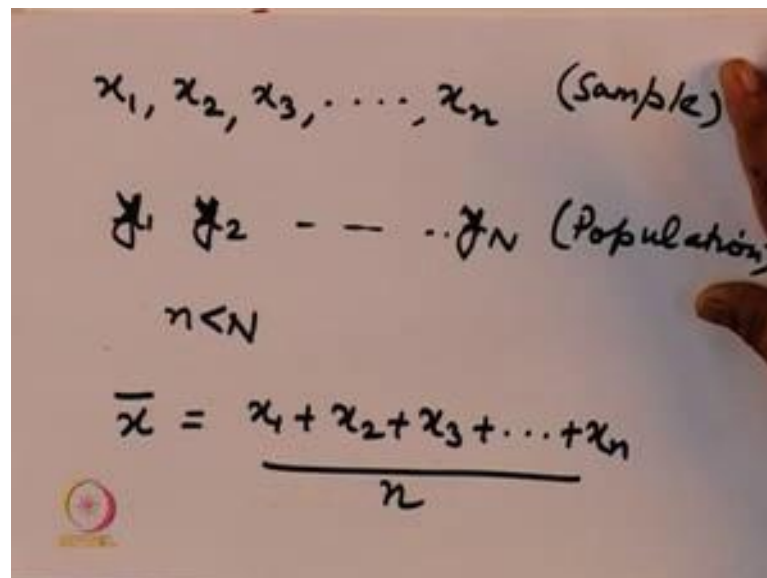
So, this you know begs the question that how can I capture this data how do I go about in trying to capture this data and that brings us to the next topic that how can I come up with numerical measures to describe data and perhaps the most popular one the widest used one is arithmetic mean which most likely all of you have heard.

(Refer Slide Time: 14:02)



So, average or arithmetic mean the way it is defined is let us say I have a set of variables $x_1, x_2, x_3, \dots, x_n$ variables x_n .

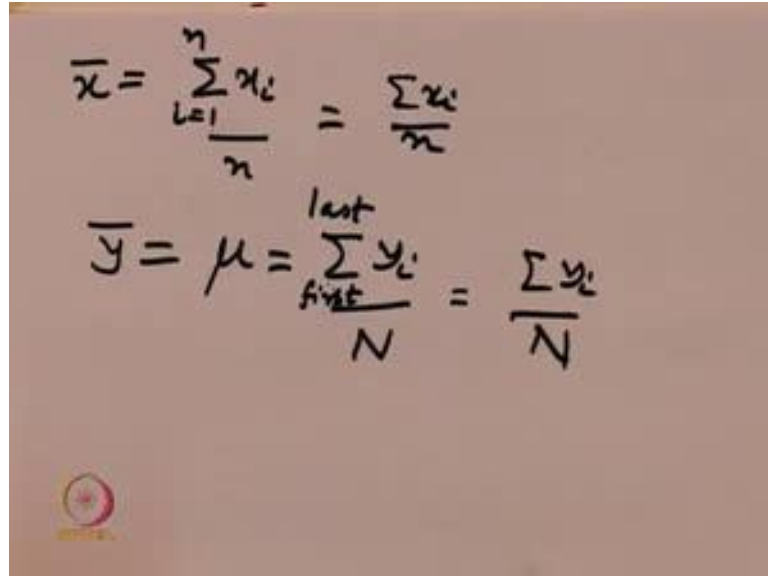
(Refer Slide Time: 14:13)



Now, n , this is my sample I have drawn a sample of n from a population y_1 or let us say $x_1, x_2, x_{\text{capital } N}$ this is my population. So, just to you know stop any stop any confusion; let me write them as y_1, y_2 and y_N . So essentially, you have derived a sample of small n ; small n is less than capital N and we want to know, what is the arithmetic average? So, as most of you are perhaps aware mean or sample mean is

defined by this variable called \bar{x} which is nothing but x_1 plus x_2 plus x_3 plus up to x_n divided by the number of observations which is n .

(Refer Slide Time: 15:17)



The image shows two handwritten mathematical formulas on a brown background. The first formula is $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum x_i}{n}$. The second formula is $\bar{y} = \mu = \frac{\sum_{\text{first}}^{\text{last}} y_i}{N} = \frac{\sum y_i}{N}$. A small logo is visible in the bottom left corner of the image.

So, in simple language in simple language I can also write it as in a compressed form I write summation x_i i is equal to one to n which means. So, you put the summation sign here and it goes from i is equal to one to i is equal to n . So, this is for i equal to one it is x_1 for i is equal to 2 it is x_2 so on and so forth. So, this is your definition of sample mean similarly, I can define my population mean. So, it is typically used μ is a term which is used to differently you know define the population mean it is similarly by y_i by capital N .

So, sometimes many times you will see we just put the summation sign without actually writing the limits, but it is understood when we do not put the limits it is from the first value to the last value. So, this is from the first value to the last value. So, in simple speak I would then. So, I can accordingly simplify \bar{x} as summation x_i by n and this is summation y_i by capital N .

(Refer Slide Time: 16:24)

The image shows a whiteboard with handwritten mathematical derivations. At the top left, the variable x is written. Below it, the equation $y = ax$ is written. To the right, the mean of y is defined as $\bar{y} = \frac{y_1 + y_2 + y_3 + \dots + y_n}{n}$. A bracket on the left groups \bar{x} and \bar{y} , with an arrow pointing to the following equations: $y_1 = ax_1$, $y_2 = ax_2$, $y_3 = ax_3$, a vertical ellipsis, and $y_n = ax_n$. To the right of these, the mean of y is expressed as $\bar{y} = \frac{ax_1 + ax_2 + \dots + ax_n}{n}$, which is then simplified to $\bar{y} = a \frac{(x_1 + x_2 + \dots + x_n)}{n}$.

So, now let us do some transformations simple thing. So, imagine I had this variable x and I define this variable y which is nothing but a times x and I want to know what is the relationship between \bar{x} and \bar{y} how are they connected. So, how do I go about it? So, what I should do is I can write, if y is equal to $a x$. So, this would mean my y_1 is equal to $a x_1$, y_2 is equal to $a x_2$, y_3 equal to $a x_3$ dot y_n is equal to $a x_n$.

So, then we can derive. So, what is my definition of \bar{y} \bar{y} has to be nothing but y_1 plus y_2 plus y_3 up to y_n by n . So, y_1 is nothing but $a x_1$ plus $a x_2$ up to $a x_n$ by n I can take a common and it is x_1 plus x_2 up to x_n by n .

(Refer Slide Time: 17:40)

The image shows a whiteboard with handwritten mathematical equations. The first equation is $\bar{y} = \frac{a(x_1 + x_2 + \dots + x_n)}{n}$. Below it, it is simplified to $= a \bar{x}$. To the right, the transformation is defined as $y = c + x$. Below that, the mean of the transformed variable is derived as $\bar{y} = \frac{(c + x_1) + (c + x_2) + \dots + (c + x_n)}{n}$. At the bottom left, the text $\bar{y} \& \bar{x}$ is written. A small logo is visible in the bottom left corner of the whiteboard.

So if this is, let me write it in a new page my \bar{y} is nothing but a times x_1 plus x_2 up to x_n by n which is nothing but \bar{x} . So, x_1 plus x_2 up to x_n by n is \bar{x} . So, \bar{y} is nothing but $a \bar{x}$. So, this tells you that when you have a pre factor which is multiplied on a variable x then the average of the new variable y is simply the pre factor multiplied by the average of x .

Now, we will let us do another simple transformation imagine y is nothing but c plus x . So, I want to ask what is \bar{y} and \bar{x} how are they related. So, what you can do the same exercise you can do the same thing, but \bar{y} is now c plus x_1 plus c plus x_2 plus c plus x_n by n .

(Refer Slide Time: 18:44)


$$\begin{aligned}\bar{y} &= \frac{c(1+1+\dots+n \text{ times}) + (x_1+x_2+\dots+x_n)}{n} \\ &= c + \bar{x} \\ y &= c + ax \\ \bar{y} &= c + a\bar{x}\end{aligned}$$

So, if that is, so what I can take in my definition of \bar{y} I can take c common, but c is 1 plus 1 plus 1 n times then you have x_1 plus x_2 plus x_n and this whole thing divided by n .

Now, this 1 plus 1 plus 1 the; you know n times is simply n . So, this gives me the formula c plus and this remaining thing is \bar{x} . So, \bar{y} is $c + \bar{x}$. So, when you have a constant to a variable x added what you can do is you can simply put \bar{y} is a constant plus the \bar{x} . So, if I were to generalize these 2 rules. So, if I have a variable y which is c plus $a x$ then I can write the formula \bar{y} is c plus $a \bar{x}$.

(Refer Slide Time: 19:40)

Transformations on Arithmetic Mean


$$y = ax$$
$$y = c + x$$
$$y = c + ax$$


(Refer Slide Time: 19:44)

Example 1: Calculating Mean using transformations

Test Scores of 20 students

61	49
93	74
87	66
42	45
55	68
67	88
82	59
50	71
29	21
55	50



So, these you know this is what that you know 3 rules of transformation is. So, let us apply this transformation to a sample example. So, imagine these are the test scores of the 20 students.

(Refer Slide Time: 19:54)

x	y
61	-4
93	+28
87	+22
42	-23
55	-10
67	2
82	17
	34

$y = x - 65$

$\bar{x} = \bar{y} + 65$

$\bar{x} \approx 70$

$\bar{y} = \frac{34}{7}$

$\approx 5 \frac{4}{7}$

So, if I write I would write down like this 61, 93, 87, 42, 55, 67, 82 so on and so forth. So, I will only take these following numbers.

So, now let us say this is my design variable x . So, I want to add the; I want to find out \bar{x} now what I see the range. So, if I just look at these few numbers this is the smallest and this is the largest. So, what I can do is if I define. So, it boils down to the question; how should I define y ? So, that it will be convenient for me to calculate this average using this using these transformations. So, what you can do one of the transformations which you might think of doing is dividing by 10. So, everything becomes 6.1, 9.3, 8.7, but still you have to add lot of the fractions which is you know not so convenient if you are doing it by hand.

So, easier alternative is to define y . So, if I define y . So, my smallest is 42, largest is 93, let us choose y as x minus some value which is in the middle of 40 and 90. So, which is 65; 65 is nothing but average of 40 and 90. So, 40 plus 90 by 2 is 65, if I define y is x minus 65 then this guy becomes minus 4, this guy becomes plus 28, this becomes plus 22, this becomes minus 23, 55 means minus 10, 67 means 2, 82 means 17. So, I can add these numbers up this is much more reasonable than adding much bigger numbers. So, what you can see is. So, 22 and 23 is almost cutting each other out. So, these 2 put together is plus one and here you have. So, 28 plus 2; 30 minus 10; 20 plus 17; 20 plus 17 plus 1; 18, 38 minus 4, so, 38 minus 4 is 34.

cancel each other out and you get a lower number. So, we want to get to as low as sum as possible that we can directly compute the total sum by hand and then accordingly we can apply this transformation.

So, this is the way to go about it you can roughly you sort the numbers from lowest to highest you identify the numbers and you roughly. So, again we did not put c exactly as 42 plus 93 by 2 because again the process would have been much more difficult to do by hand we wanted an approximation which we can which is easy to implement and we can see by hand what we are doing.

(Refer Slide Time: 25:11)

Example 2: Calculating Mean using transformations

Body Mass Index (BMI)

18.3 21.9 23.0 24.3 25.4

26.6 27.5 28.8 34.2 31.0

19.2 21.0 24.5 25.5 27.8

28.2 31.0 29.1 28.1 24.2

25.6 20.0 20.0 25.0 25.2

The slide displays a list of 20 BMI values arranged in five rows of five. The values are: 18.3, 21.9, 23.0, 24.3, 25.4; 26.6, 27.5, 28.8, 34.2, 31.0; 19.2, 21.0, 24.5, 25.5, 27.8; 28.2, 31.0, 29.1, 28.1, 24.2; 25.6, 20.0, 20.0, 25.0, 25.2. A small logo is visible in the bottom left corner of the slide.

So, let us just take one more example which is the next example. So, again you see that you have you have this entire range of numbers which goes from. So, this is body mass indices it goes. So, if I look at look through the screen my lowest the smallest number I have is probably 18 by 3; 18.3. So, smallest number is 18.3.

(Refer Slide Time: 25:31)

Handwritten calculations on a whiteboard:

$$\begin{aligned} \text{Smallest} &= 18.3 & 18.3 - 24 &= -6 \\ \text{Largest} &= 29.1 & 29.1 - 24 &= 5 \end{aligned}$$
$$\begin{aligned} \text{Av.} &= \frac{18.3 + 29.1}{2} \\ &= \frac{47.4}{2} \approx 24 \end{aligned}$$

So, this is my smallest the largest number, largest number is 29.1; 29.1. So, then my average is going to be 18.3 plus 29.1 by 2. So, 29 and 18 is 47; 47.4 by 2 is roughly equal to 24. So, it is 24.

So, your average is roughly 24. So, you can then add it up. So, if I look at this score again. So, my 18.3 will give me a value of 18.3 minus 24 while this 29.1; 29.1 will lead minus 24 will give me roughly 5 and this will roughly minus 6. So, you see that these 2 numbers if I add up then I will get a value which is close to 0. In this way we can make use of transformations to make our life lot easier particularly when we are not working with the calculator. So, with that I would stop in this lecture.

So, summary I would summarize by saying that arithmetic mean represents one of the simplest ways of representing data odds extracting some quantitative metric to characterize your data and then using basic transformations you can simplify and do things by hand to find out what is the arithmetic mean of your given data set. So, either y is equal to c plus x or y is equal to you know c x or y is equal to a plus c x are 3 simple transformations which you can do. So, imagine if your numbers were 100, 200 like that then by 100 dividing by 100 and making it 1, 2, 3, would be one simple way.

With that I thank you for your cooperation. And I look forward to meeting you in the next lecture.

Thank you.