INDIAN INSTITUTE OF TECHNOLOGY MADRAS

सिद्धिर्भवति कर्मजा

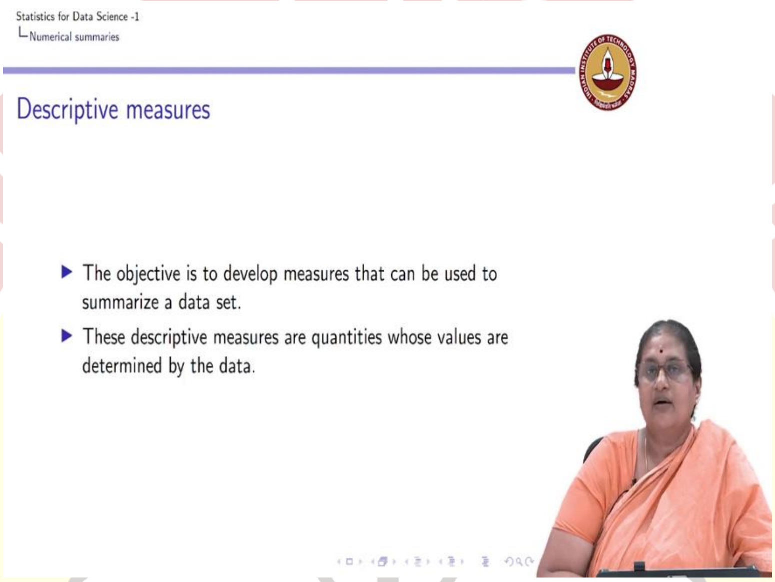
# IIT Madras

ONLINE DEGREE

**Statistics for Data Science - 1**
**Prof. Usha Mohan**
**Department of Management Studies**
**Indian Institute of Technology, Madras**

**Lecture – 3.2**
**Describing Numerical Data – Mean**

(Refer Slide Time: 00:14)



Statistics for Data Science -1
└Numerical summaries

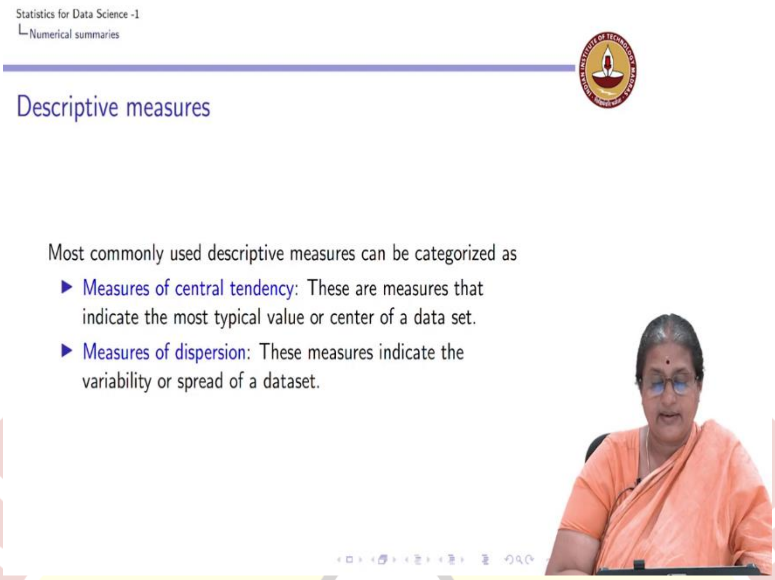
Descriptive measures

- The objective is to develop measures that can be used to summarize a data set.
- These descriptive measures are quantities whose values are determined by the data.

The next thing which we are going to discuss is how do we summarize data using numerical summaries. When we talked about the categorical variable, we introduce two descriptive measures; namely, the mode and the median. While mode was used both for nominal and ordinal type of data. We saw that if you want to summarize the data using median, it has to be a ordinal data or there should be some sense of order in the data.

So, now what we are going to look at is what are the ways we can describe numerical data. So, the objective here is to actually develop measures or summaries of data which can be used to summarize a complete data set.

(Refer Slide Time: 01:07)



Statistics for Data Science -1
└Numerical summaries

Descriptive measures

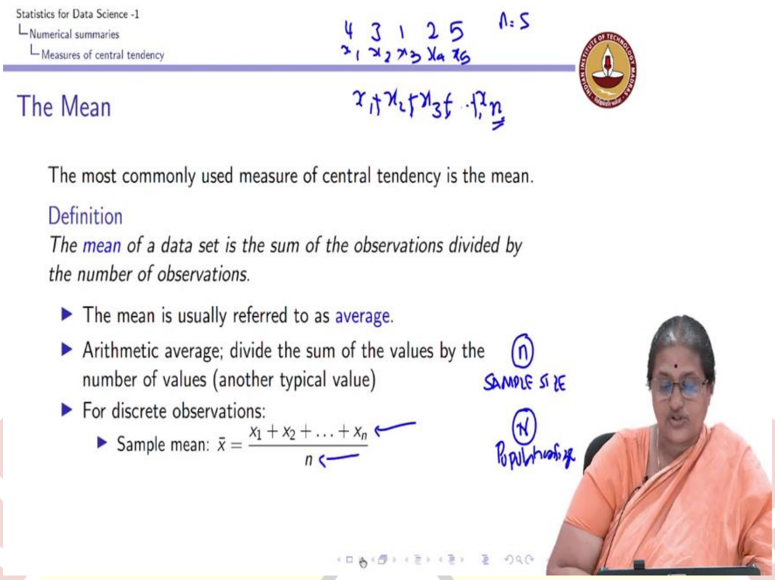
Most commonly used descriptive measures can be categorized as

- Measures of central tendency: These are measures that indicate the most typical value or center of a data set.
- Measures of dispersion: These measures indicate the variability or spread of a dataset.

So, the most commonly used, so especially when we are dealing with numerical data the most commonly used descriptive measures again can be broadly categorized into two categories and these categories are one which are known as measures of central tendency and the second which are known as measures of dispersion.

Now, what do we mean by measures of central tendency? Measures of central tendency as the name suggests, it talks about where the data is actually concentrated or what is the most typical value of a data set. A measure of central tendency describes or tells us what can we expect as a typical value of a data set; whereas, the measure of dispersion also referred to as measures of variation or measures or spread talk about the variability in a data set or spread in a data set.

(Refer Slide Time: 02:15)



Statistics for Data Science -1
└ Numerical summaries
└ Measures of central tendency

The Mean

The most commonly used measure of central tendency is the mean.

Definition

*The mean of a data set is the sum of the observations divided by the number of observations.*

- The mean is usually referred to as average.
- Arithmetic average; divide the sum of the values by the number of values (another typical value)
- For discrete observations:
  - Sample mean: $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$

4 3 1 2 5
n:5
$x_1$ $x_2$ $x_3$ $x_4$ $x_5$
$x_1 + x_2 + x_3 + \cdots + x_n$
n
SAMPLE SIZE
N
Population size

So, we will go through these measures in detail. When we talk about measures of central tendency, the most common measure of central tendency is what we refer to as the mean of a data set.

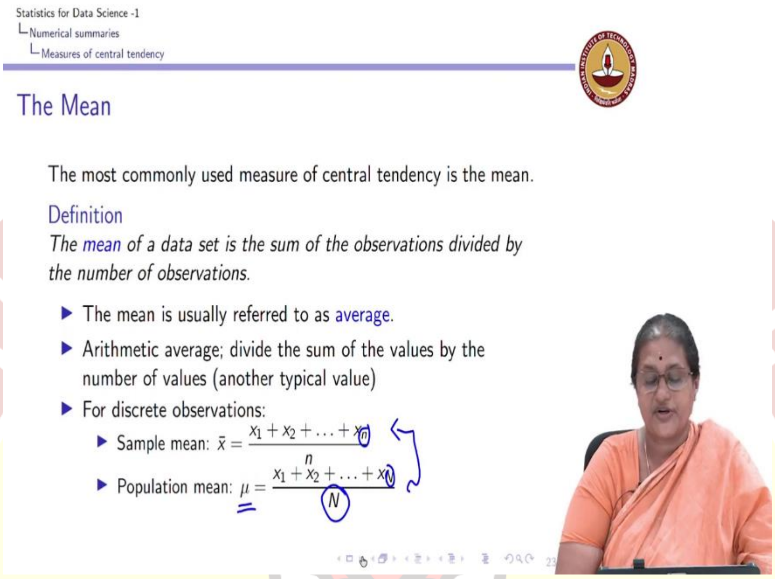
What is a mean? I define the mean of a data set to be the sum of the observations divided by the number of observations. So, now, let us formally define what it is. Suppose, I have $x_1$, $x_2, x_3$, I refer to my observations as $x_1$, $x_2, x_3$,. For example, I have my data which is 4, 3, 1, 2 and 5. I have $n$ equal to 5 here. My $x_1$ is 4; $x_2$is 3; $x$_3 is 1; $x$_4 is 2 and $x$_5 is 0.

So, in general, if I have $n$ observations each $x$_1 refers to the first observation in my data set; $x$_2 refers to the second observation in my data set and $x$_$n$ refers to the $nth$ observation in my data set. The mean of a data set is the sum of these observations, the numerator gives me the sum of these observations which is nothing but $x_1 + x_2 + x_3 \ldots +$ $x$_$n$ divided by the total number of observations which is $n$, small $n$.

Now, recall in when we introduce the notion of a sample and population, we said that a small $n$ refers to what we call the sample size; whereas, capital $N$ refers to a population size. So, if I am having a data set which is a sample, then I do denote it or my notation is going to be small $n$; when I have population, when I refer to it as a population, it is going to be capital $N$.

So, if my I can define a sample mean, when I have a sample data set to be the sum of the sample observations divided by the total number of observations, the way we have done here.

(Refer Slide Time: 04:46)



Statistics for Data Science -1
└Numerical summaries
└Measures of central tendency

The Mean

The most commonly used measure of central tendency is the mean.

Definition

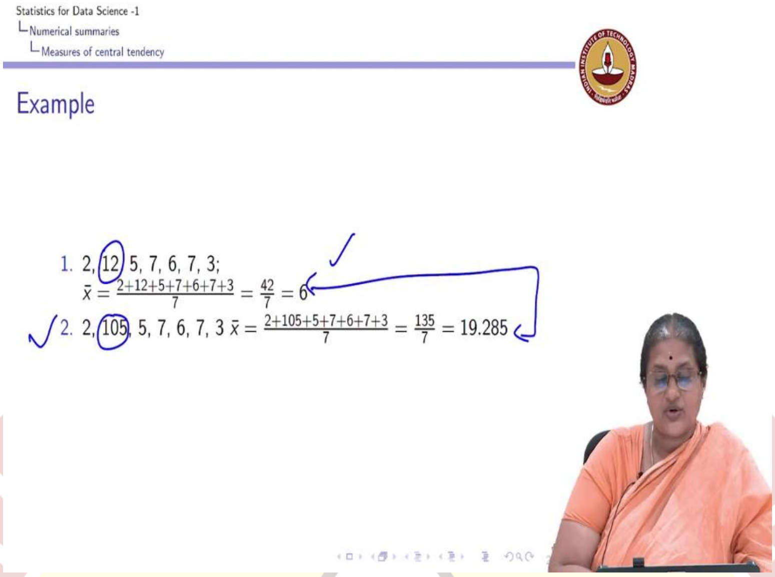
*The mean of a data set is the sum of the observations divided by the number of observations.*

- The mean is usually referred to as average.
- Arithmetic average; divide the sum of the values by the number of values (another typical value)
- For discrete observations:
  - Sample mean: $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$
  - Population mean: $\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$

I can also define a population mean which is typically refer to by the Greek alphabet μto be $x_1 + x_2 \dots + x_n$; whereas, again remember, I said $n$ is the total number of observations in a population divided by capital $N$ which is the total number of observations.

The definitions are the same, only thing the number of observations differ whether you are referring to a sample or whether you are referring to a population. The mean is also popularly refer to as a average. So, now, let us compute the mean for small data sets.

(Refer Slide Time: 05:26)



Statistics for Data Science -1
└Numerical summaries
└Measures of central tendency

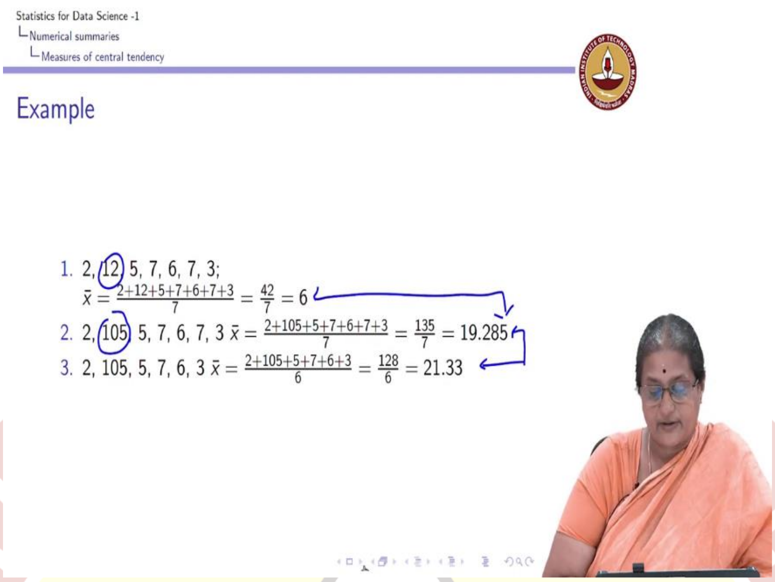
Example

1. 2, 12, 5, 7, 6, 7, 3;
   $\bar{x} = \frac{2+12+5+7+6+7+3}{7} = \frac{42}{7} = 6$
2. 2, 105, 5, 7, 6, 7, 3 $\bar{x} = \frac{2+105+5+7+6+7+3}{7} = \frac{135}{7} = 19.285$

Now, I have this data set which is given to me. So, the mean for this data set is very simple. That is $\bar{x}$ is $2 + 12 + 5 + 7 + 6 + 7 + 3$. This is what I have here and that I can see is equal to $42$, the numerator and which is equal to $6$. In other words, $6$ is the mean of this data set. Now, let us look at another data set.

See the only difference between the first data set and the second data set is in $x_2$ or the second observation, everything else is the same. So, I compute the mean for this data set, When I compute the mean for this data set, I find $\bar{x}$ is equal to $135$ by $7$ which is $19.285$. What I want you all to observe very clearly is the difference between this data set and this data set is only the second number; but the mean difference is very high.

(Refer Slide Time: 06:52)



Statistics for Data Science -1
└Numerical summaries
└Measures of central tendency

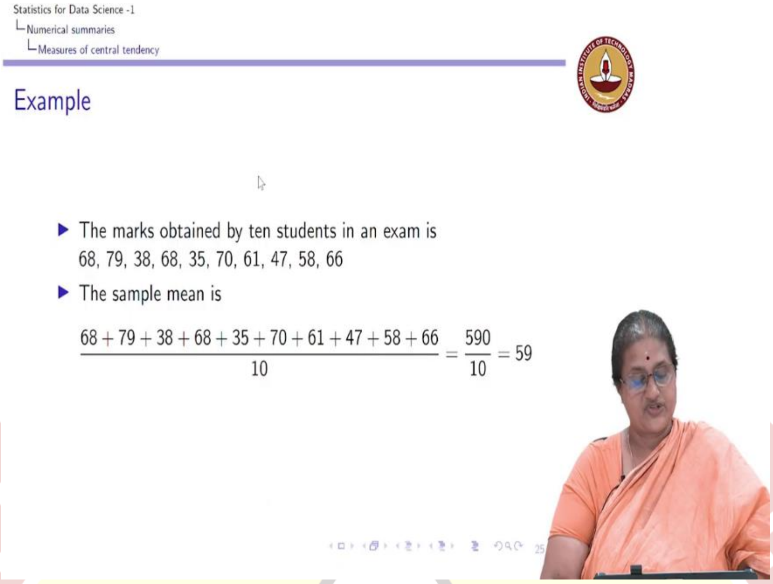
## Example

1. 2, 12, 5, 7, 6, 7, 3; $\bar{x} = \frac{2+12+5+7+6+7+3}{7} = \frac{42}{7} = 6$
2. 2, 105, 5, 7, 6, 7, 3 $\bar{x} = \frac{2+105+5+7+6+7+3}{7} = \frac{135}{7} = 19.285$
3. 2, 105, 5, 7, 6, 3 $\bar{x} = \frac{2+105+5+7+6+3}{6} = \frac{128}{6} = 21.33$

Now, let us look at another data set which is the following, where I the it has only 6 observations now, it does not have the last observation. The mean for this data set is going to be 21.33. We see that these two means are fairly close to each other; whereas, these two means are very different from each other.

So, what this tells us is the mean even though here I have only one observation which is different. This observation and this observation are very different from each other. So, the mean is very sensitive to outliers. By outliers, I mean what is the number which is very different from what the typical data set behaves like.

(Refer Slide Time: 07:48)



Statistics for Data Science -1
└Numerical summaries
└Measures of central tendency

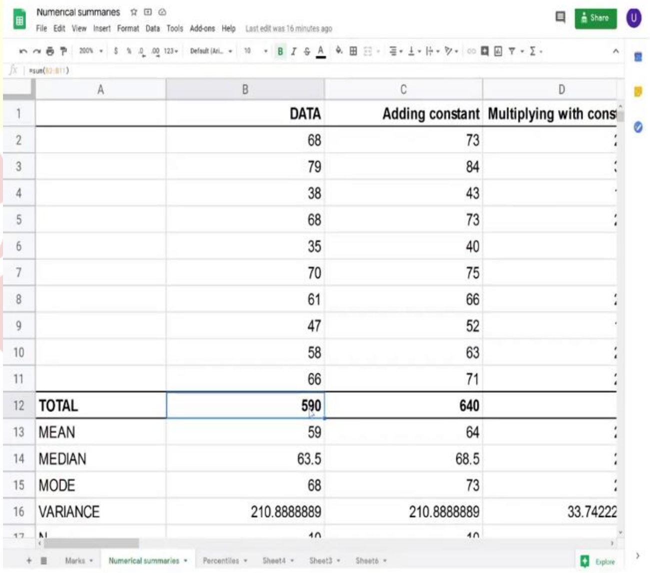
Example

- The marks obtained by ten students in an exam is 68, 79, 38, 68, 35, 70, 61, 47, 58, 66
- The sample mean is

$$\frac{68+79+38+68+35+70+61+47+58+66}{10} = \frac{590}{10} = 59$$

So, now let us go back to another example. These are the marks obtained by 10 students in an exam. So, now, I can see that the sample mean in this case is 590 divided by 10 which is 59.

(Refer Slide Time: 08:07)



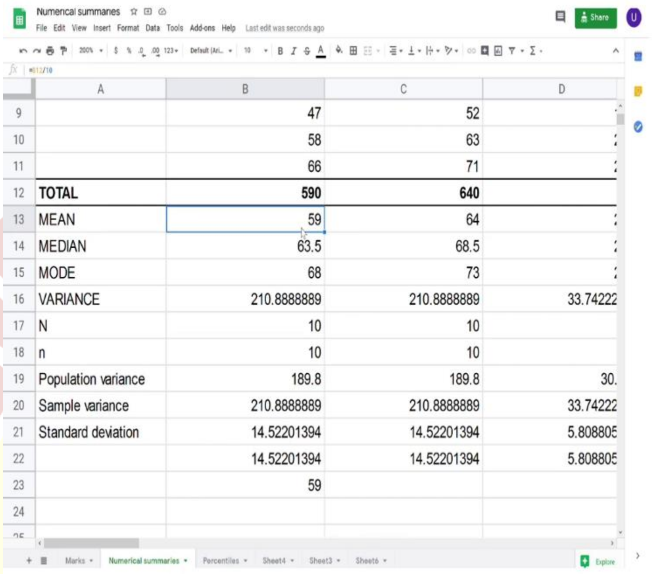
=sum(B2:B11)

| | A | B | C | D |
|---|---|---|---|---|
| 1 | | DATA | Adding constant | Multiplying with cons |
| 2 | | 68 | 73 | |
| 3 | | 79 | 84 | |
| 4 | | 38 | 43 | |
| 5 | | 68 | 73 | |
| 6 | | 35 | 40 | |
| 7 | | 70 | 75 | |
| 8 | | 61 | 66 | |
| 9 | | 47 | 52 | |
| 10 | | 58 | 63 | |
| 11 | | 66 | 71 | |
| 12 | TOTAL | 590 | 640 | |
| 13 | MEAN | 59 | 64 | |
| 14 | MEDIAN | 63.5 | 68.5 | |
| 15 | MODE | 68 | 73 | |
| 16 | VARIANCE | 210.8888889 | 210.8888889 | 33.74222 |
| 17 | N | 10 | 10 | |

So, this is the data I have. So, these are the 10 data sets. This is the data which I have, which I have just given to you $68, 79$; this is the data which we have worked with. So, in this data, you can see that this sum $B2$ to $B11$ is nothing but the sum which I have worked here.

So, the sum of this is the numerator sum which is equal to 590. I have a 590 here which is the sum of all my data values and you can see the mean, the google sheet the mean gives me what is $B12$ by 10 that is what is 590 divided by the total number of observations.

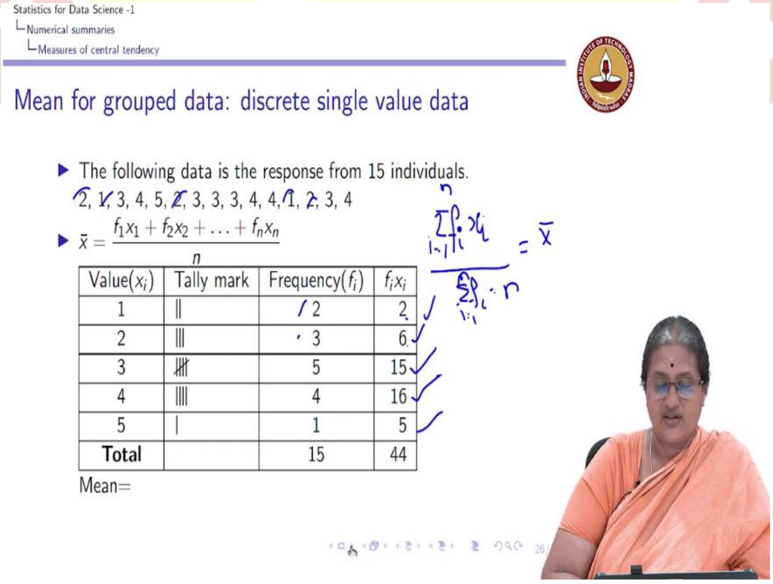
(Refer Slide Time: 09:06)



|  | A | B | C | D |
|---|---|---|---|---|
| 9 |  | 47 | 52 |  |
| 10 |  | 58 | 63 |  |
| 11 |  | 66 | 71 |  |
| 12 | TOTAL | 590 | 640 |  |
| 13 | MEAN | 59 | 64 |  |
| 14 | MEDIAN | 63.5 | 68.5 |  |
| 15 | MODE | 68 | 73 |  |
| 16 | VARIANCE | 210.8888889 | 210.8888889 | 33.74222 |
| 17 | N | 10 | 10 |  |
| 18 | n | 10 | 10 |  |
| 19 | Population variance | 189.8 | 189.8 | 30. |
| 20 | Sample variance | 210.8888889 | 210.8888889 | 33.74222 |
| 21 | Standard deviation | 14.52201394 | 14.52201394 | 5.808805 |
| 22 |  | 14.52201394 | 14.52201394 | 5.808805 |
| 23 |  | 59 |  |  |
| 24 |  |  |  |  |

That is what the mean gives you. Now, you can see that the google sheet has average demand, command is giving me the same average as what I have obtained before. So, average command in the google sheet gives me the mean which we have different ok. The next thing which we are going to see is how do I obtain the mean for a grouped data.

(Refer Slide Time: 09:36)



Statistics for Data Science -1
└Numerical summaries
  └Measures of central tendency

Mean for grouped data: discrete single value data

- The following data is the response from 15 individuals.
  2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4
- $\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{n}$

| Value($x_i$) | Tally mark | Frequency($f_i$) | $f_ix_i$ |
|---|---|---|---|
| 1 | \|\| | 2 | 2 |
| 2 | \|\|\| | 3 | 6 |
| 3 | ~~\|\|\|\|~~ | 5 | 15 |
| 4 | \|\|\|\| | 4 | 16 |
| 5 | \| | 1 | 5 |
| **Total** |  | 15 | 44 |

Mean=

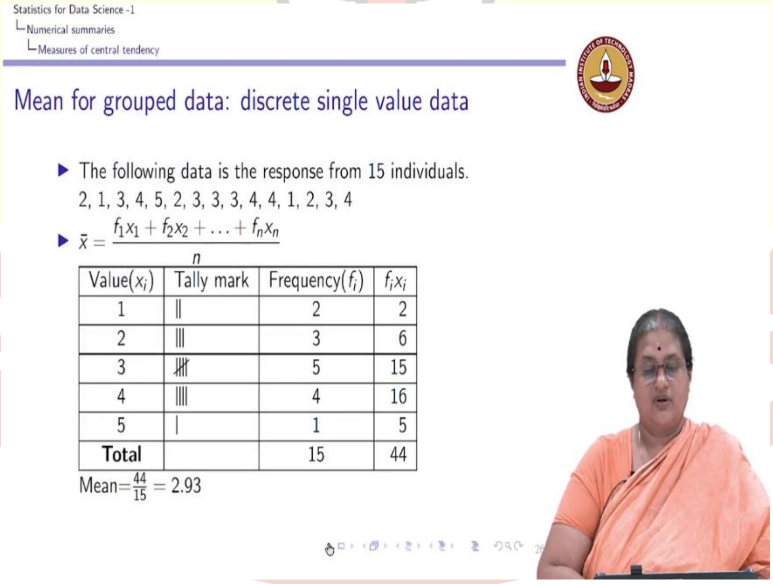
$$\frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i = n} = \bar{x}$$

Now, if I have single. So, now, let us go back to this data which we have seen earlier which was the response from 15 individuals, we have already seen that this is the frequency 1 appears twice; 2 appears thrice; 3 appears five times; 4 appears four times and 5 appears once, this is what we have already seen.

Now, for a data of this kind how do I compute the average? Now, 1 appears two times, so it appears with the frequency 2. So, the way I need to add 1 2 times. So, it is $2 \times 1$ so $1 + 1$ which is giving me a 2. Similarly, 2 appears 3 times. So, the total sum of 2 is going to be $2 + 2 + 2$ or $2 \ \times 3$ which is equal to 6. 3 appears five times, so the sum 3 contributes to its $3 \ \times 5$ which is 15; 4 into 4 which is 16 and 5 into 1 is 15.

So, instead of writing $2 + 1 + 3 +$ all of it, I can write it as 1 appears two times which contributes to 2; 2 appears three times which contributes six and this is the numerator which is $\sum_{i=1}^{n} f_i x_i$ divided by your $n$ or $\sum f_i$ which is equal to $n$. This gives me the $\bar{x}$ when I have discrete single value data.

(Refer Slide Time: 11:31)



Statistics for Data Science -1
└Numerical summaries
 └Measures of central tendency

Mean for grouped data: discrete single value data

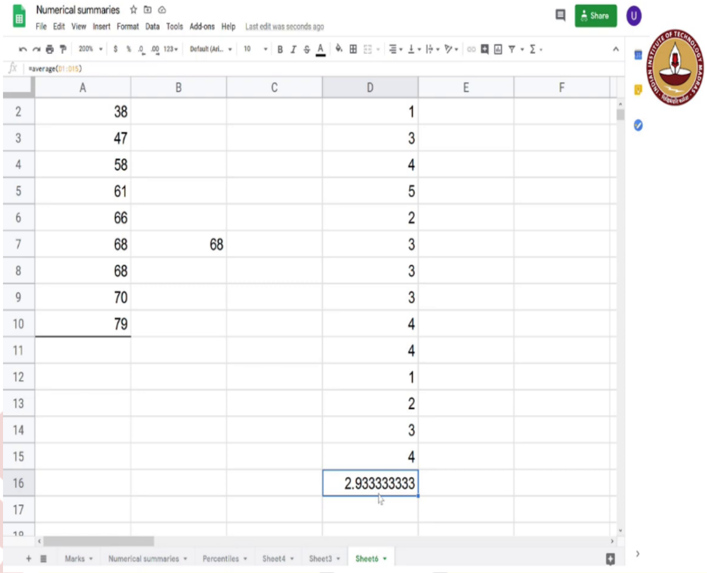
- The following data is the response from 15 individuals.
  2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4
- $\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{n}$

| Value($x_i$) | Tally mark | Frequency($f_i$) | $f_ix_i$ |
|---|---|---|---|
| 1 | \|\| | 2 | 2 |
| 2 | \|\|\| | 3 | 6 |
| 3 | 𝍸 | 5 | 15 |
| 4 | \|\|\|\| | 4 | 16 |
| 5 | \| | 1 | 5 |
| **Total** | | 15 | 44 |

Mean$=\frac{44}{15} = 2.93$

So, the mean here is going to be 44 divided by 15 which is equal to 2.93.

(Refer Slide Time: 11:55)



Numerical summaries

File Edit View Insert Format Data Tools Add-ons Help Last edit was seconds ago

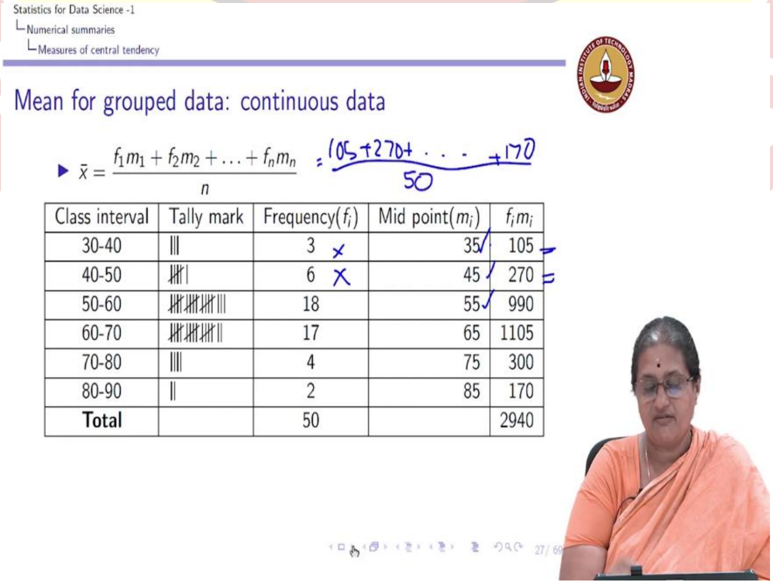
=average(D1:D15)

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 2 | 38 | | | 1 | | |
| 3 | 47 | | | 3 | | |
| 4 | 58 | | | 4 | | |
| 5 | 61 | | | 5 | | |
| 6 | 66 | | | 2 | | |
| 7 | 68 | 68 | | 3 | | |
| 8 | 68 | | | 3 | | |
| 9 | 70 | | | 3 | | |
| 10 | 79 | | | 4 | | |
| 11 | | | | 4 | | |
| 12 | | | | 1 | | |
| 13 | | | | 2 | | |
| 14 | | | | 3 | | |
| 15 | | | | 4 | | |
| 16 | | | | 2.933333333 | | |
| 17 | | | | | | |

Marks · Numerical summaries · Percentiles · Sheet4 · Sheet3 · Sheet6

Now, if I go back and type in the same data in a particular sheet and I go back and type in the same data in a particular sheet, I add a sheet, I just type in this data. So, the data is $1, 2, 1, 3, 4, 5.$ So, I am just going to type in the data. $2, 1, 3, 4, 5. 2, 3, 3, 3, 4;\ 2, 3, 3, 3, 4\ 4, 1, 2, 3, 4;\ 4, 1, 2, 3, 4.$ I can see that there are 15 values, I just write down the average of this data set which you can see is equal to 2.93 and this is precisely what we had got in our earlier.

(Refer Slide Time: 13:01)



Statistics for Data Science -1
└Numerical summaries
 └Measures of central tendency

Mean for grouped data: continuous data

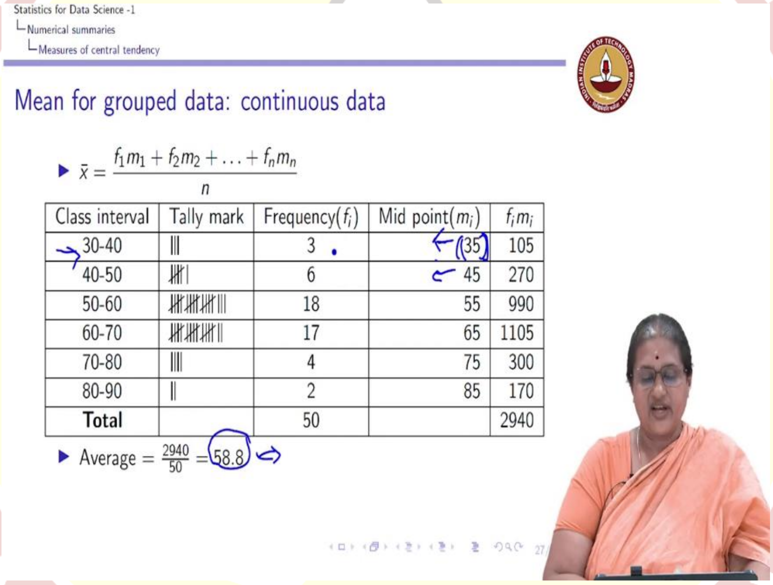
- $\bar{x} = \dfrac{f_1m_1 + f_2m_2 + \ldots + f_nm_n}{n} = \dfrac{105 + 270 + \cdot\cdot\cdot + 170}{50}$

| Class interval | Tally mark | Frequency($f_i$) | Mid point($m_i$) | $f_im_i$ |
|---|---|---|---|---|
| 30-40 | \|\|\| | 3 | 35 | 105 |
| 40-50 | 𝍸\| | 6 | 45 | 270 |
| 50-60 | 𝍸𝍸𝍸\|\|\| | 18 | 55 | 990 |
| 60-70 | 𝍸𝍸𝍸\|\| | 17 | 65 | 1105 |
| 70-80 | \|\|\|\| | 4 | 75 | 300 |
| 80-90 | \|\| | 2 | 85 | 170 |
| **Total** | | 50 | | 2940 |

Now, what how do we compute the mean for group data or continuous data? Now, again recall this when I have continuous data, I do not have a particular value of the discrete data that is being taken. So, in this case recall we have what is called the midpoint of the data set or for each class interval. For example, the midpoint of the class interval 30 to 40 is 35; 40 to 50 is 45; 50 to 60 is 55. So, you look at the midpoint rather than looking at each discrete value, I know the frequency of the data set.

Now, I multiply this frequency with the midpoint, I get 105 and 6 into 45 is 270. My numerator is going to be 105 plus 270 plus 170 divided by the total number is again 50. So, it is 2940 divided by 50 which is 58.8.

(Refer Slide Time: 14:11)



Statistics for Data Science -1
└Numerical summaries
└Measures of central tendency

Mean for grouped data: continuous data

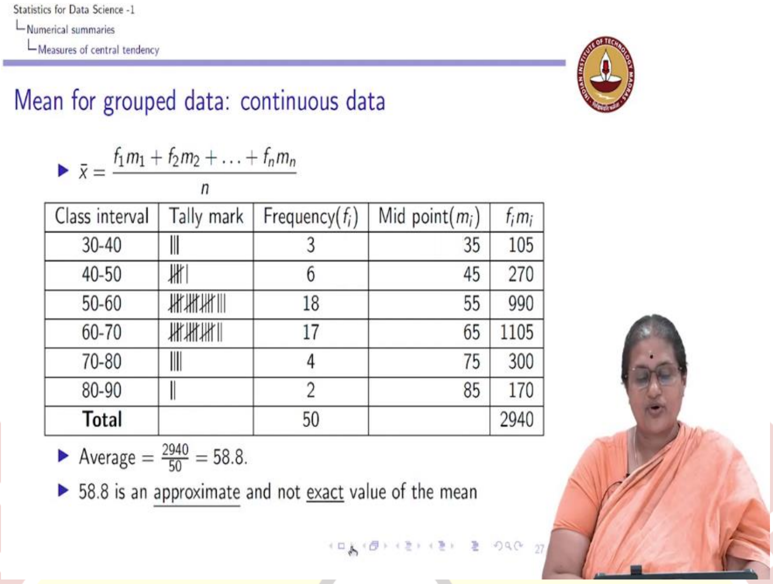
▶ $\bar{x} = \frac{f_1m_1 + f_2m_2 + \ldots + f_nm_n}{n}$

| Class interval | Tally mark | Frequency($f_i$) | Mid point($m_i$) | $f_im_i$ |
|---|---|---|---|---|
| 30-40 | \|\|\| | 3 | 35 | 105 |
| 40-50 | 卌\| | 6 | 45 | 270 |
| 50-60 | 卌卌卌\|\|\| | 18 | 55 | 990 |
| 60-70 | 卌卌卌\|\| | 17 | 65 | 1105 |
| 70-80 | \|\|\|\| | 4 | 75 | 300 |
| 80-90 | \|\| | 2 | 85 | 170 |
| **Total** | | 50 | | 2940 |

▶ Average = $\frac{2940}{50}$ = 58.8

A word of caution here is this 58.8 is not the actual mean because we are approximating it only with the midpoint, I am now taking the actual values of the data. I repeat this 58.8 is an approximation because I am multiplying the frequency with the midpoint or the best representative in this class interval which is 35. I know my data lies between 30 to 40. So, I am just saying that ok, it is around 35. So, this 35 is an approximation.

(Refer Slide Time: 15:04)



Statistics for Data Science -1
└Numerical summaries
  └Measures of central tendency

## Mean for grouped data: continuous data

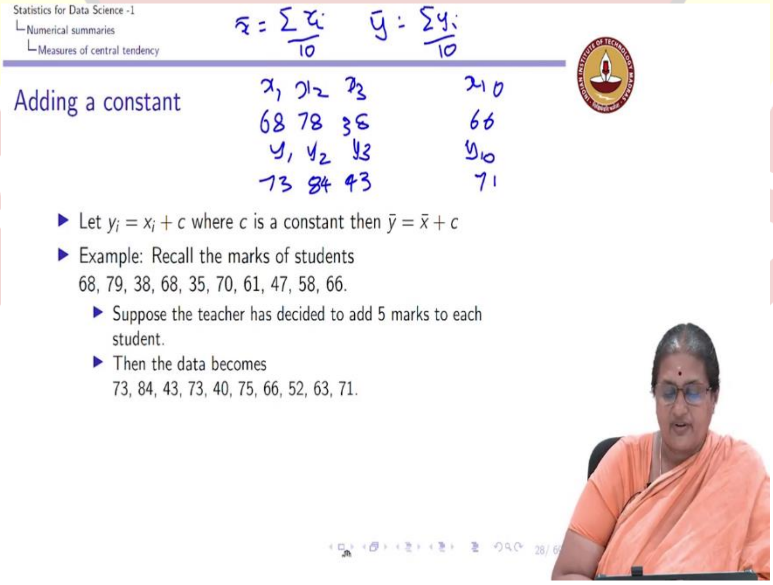
- $\bar{x} = \frac{f_1 m_1 + f_2 m_2 + \ldots + f_n m_n}{n}$

| Class interval | Tally mark | Frequency($f_i$) | Mid point($m_i$) | $f_i m_i$ |
|---|---|---|---|---|
| 30-40 | \|\|\| | 3 | 35 | 105 |
| 40-50 | 𝍸\| | 6 | 45 | 270 |
| 50-60 | 𝍸𝍸𝍸\|\|\| | 18 | 55 | 990 |
| 60-70 | 𝍸𝍸𝍸\|\| | 17 | 65 | 1105 |
| 70-80 | \|\|\|\| | 4 | 75 | 300 |
| 80-90 | \|\| | 2 | 85 | 170 |
| **Total** | | 50 | | 2940 |

- Average $= \frac{2940}{50} = 58.8$.
- 58.8 is an approximate and not exact value of the mean

So, because we are not looking at the exact data values, this average is only an approximation and not the exact value of the mean; whereas, when we looked at the discrete single value data, I took the exact value of my data and hence, my mean matched with the exact average.

(Refer Slide Time: 15:24)



Statistics for Data Science -1
└Numerical summaries
  └Measures of central tendency

## Adding a constant

$\bar{x} = \frac{\sum x_i}{10}$ $\bar{y} = \frac{\sum y_i}{10}$

$x_1$ $x_2$ $x_3$ $x_{10}$
68 78 38 66
$y_1$ $y_2$ $y_3$ $y_{10}$
73 84 43 71

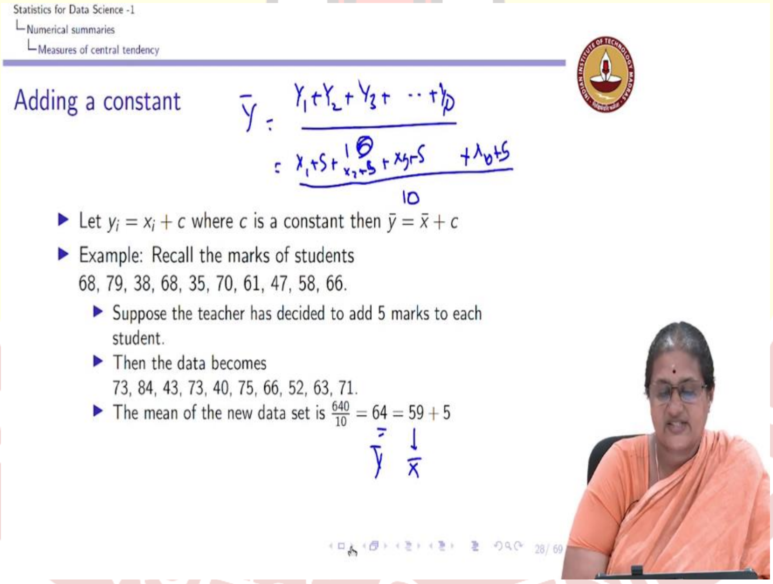
- Let $y_i = x_i + c$ where $c$ is a constant then $\bar{y} = \bar{x} + c$
- Example: Recall the marks of students
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66.
  - Suppose the teacher has decided to add 5 marks to each student.
  - Then the data becomes
    73, 84, 43, 73, 40, 75, 66, 52, 63, 71.

This is a point to be noted. Now, how sensitive is the mean if I add a constant to every observation. Now, why are we even interested in knowing about this? For example, let us look at a case where a teacher has already given the marks to the students that is I have marks of 10 students. Now, I become benevolent, suddenly and I decide to add 5 marks to each of the students.

So, this is the earlier data set which was 68 had a mean of 59. Now, I am adding 5 marks to each of these students. So, now, the data set becomes 68 +5 which is 73; 79 + 5 which is 84 and so forth 66 + 5 which is 71. The number of observations remain the same, but I am adding a constant that is I am adding 5 to each value in my data set ok. So, what happens to the mean?

So, earlier my $x_1, x_2, x_n$ were $68, 78, x_3$ was 38, my $x_{10}$ was 66. I am going to add $y_1$ is become 73, 68 +5; $y_2$ is 84 again 78 + 5; $y_3$ is 43, 38 +5; $y_{10}$ is 71. So, by my definition, if I define what is $\bar{x}, \bar{x}$ is going to be a $\frac{\sum x_i}{10}$; $\bar{y}$ is going to be $\frac{\sum y_i}{10}$. This is what our definitions say.

(Refer Slide Time: 17:35)



Statistics for Data Science -1
└Numerical summaries
 └Measures of central tendency

Adding a constant

$\bar{y} = \frac{y_1+y_2+y_3+\cdots+y_{10}}{}$

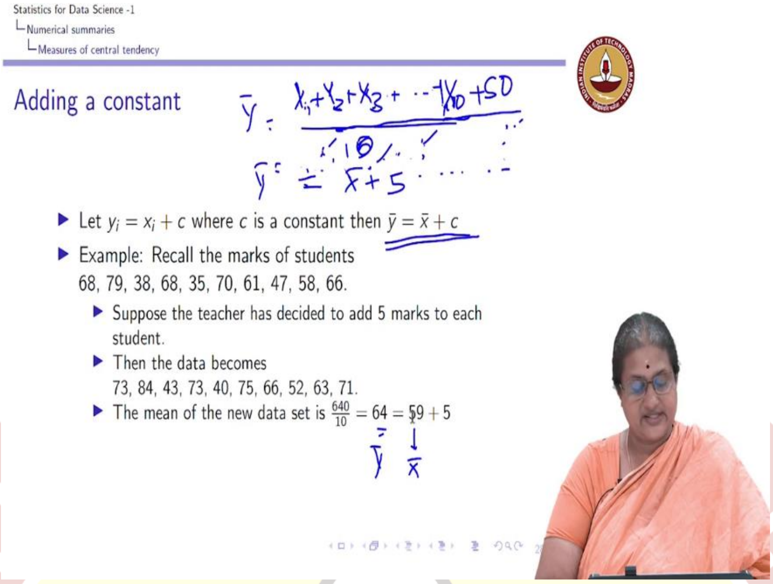
$= \frac{x_1+5+x_2+5+x_3+5 \quad +x_{10}+5}{10}$

- Let $y_i = x_i + c$ where $c$ is a constant then $\bar{y} = \bar{x} + c$
- Example: Recall the marks of students
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66.
  - Suppose the teacher has decided to add 5 marks to each student.
  - Then the data becomes
    73, 84, 43, 73, 40, 75, 66, 52, 63, 71.
  - The mean of the new data set is $\frac{640}{10} = 64 = 59 + 5$

$\underset{\bar{y}}{\overset{}{=}} \quad \downarrow \bar{x}$

So, if I add this and I compute $y_1$, I find out the $y_1$ is going to be $\frac{640}{10}$; $\frac{640}{10}$ which is 64. You can notice that 64 is my $\bar{y}$. This is my $\bar{x}$which is 5. So, why does this happen? So, I know that $\frac{y_1+y_2+y_3+\cdots+y_{10}}{5\ or\ 10}$ is my $\bar{y}$; but $y_1$ is $x_1 + 5$, $y_2$ is $x_2 + 5, y_3$ is $x_3 + 5$. Similarly, $y_{10}$ is also $x_{10} + 5$. So, you can see that I have $x_1 + x_2 + x_3 + x_{10}$ which makes $\bar{y}$ equal to the following.

(Refer Slide Time: 18:35)



Statistics for Data Science -1
└Numerical summaries
└Measures of central tendency

Adding a constant

$$\bar{y} = \frac{x_1+y_2+x_3+\cdots+x_{10}+50}{10}$$

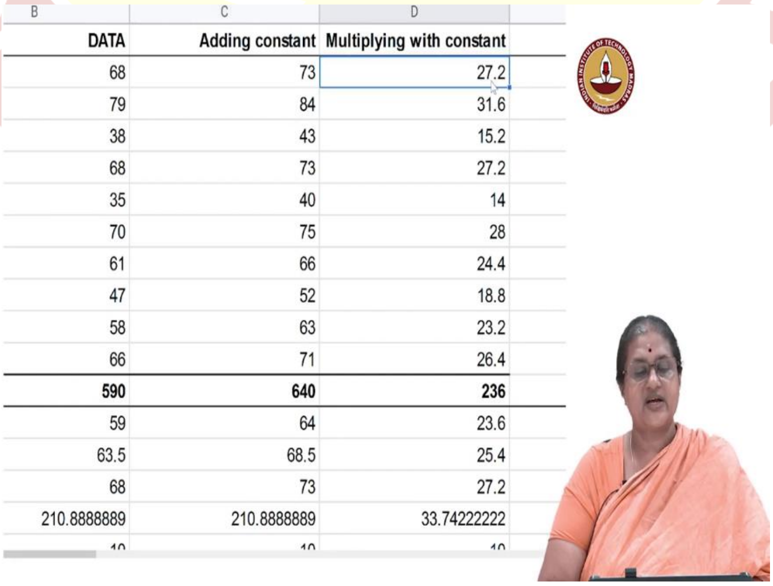
$$\bar{y} = \bar{x}+5$$

- Let $y_i = x_i + c$ where $c$ is a constant then $\underline{\underline{\bar{y} = \bar{x} + c}}$
- Example: Recall the marks of students 68, 79, 38, 68, 35, 70, 61, 47, 58, 66.
  - Suppose the teacher has decided to add 5 marks to each student.
  - Then the data becomes 73, 84, 43, 73, 40, 75, 66, 52, 63, 71.
  - The mean of the new data set is $\frac{640}{10} = \underset{\bar{y}}{64} = \underset{\bar{x}}{59} + 5$

So, I have this which is nothing but $\frac{x_1+x_2+x_3+x_{10}+50}{10}$ because this 5 is added 10 times. Now, $\frac{x_1+x_2+x_3+x_{10}}{10}$ is my $\bar{x}$ that is what we have already seen. $x_1 + x_2 + x_3 + x_{10}$ is my $\bar{x}$ that is something which we have already seen ok.

So, I have this as my $\bar{x}$ $50 + 10$ is 5, this is the constant. So, I have $\bar{y}$ which is equal to $\bar{x} + c$. So, in summary, when I add a constant to every data point what happens is the mean of the new data set is the old mean plus the same constant.

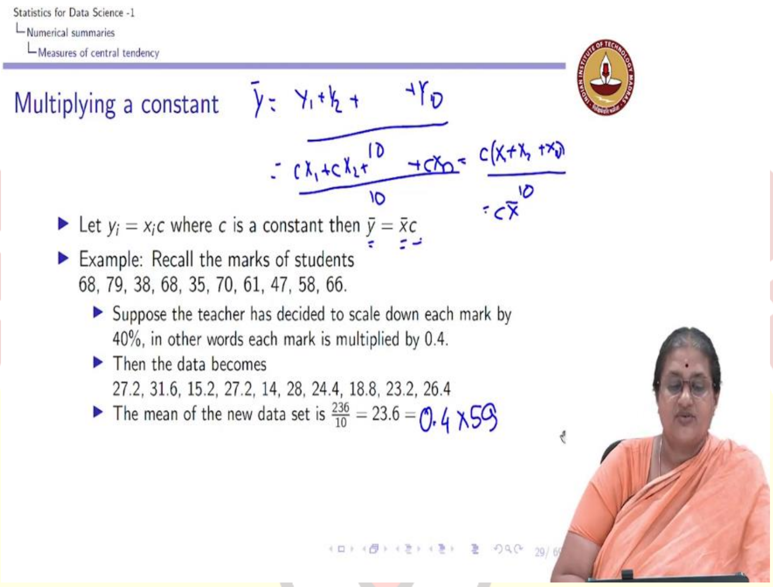
(Refer Slide Time: 19:42)



| B | C | D |
|---|---|---|
| DATA | Adding constant | Multiplying with constant |
| 68 | 73 | 27.2 |
| 79 | 84 | 31.6 |
| 38 | 43 | 15.2 |
| 68 | 73 | 27.2 |
| 35 | 40 | 14 |
| 70 | 75 | 28 |
| 61 | 66 | 24.4 |
| 47 | 52 | 18.8 |
| 58 | 63 | 23.2 |
| 66 | 71 | 26.4 |
| **590** | **640** | **236** |
| 59 | 64 | 23.6 |
| 63.5 | 68.5 | 25.4 |
| 68 | 73 | 27.2 |
| 210.8888889 | 210.8888889 | 33.74222222 |
| 10 | 10 | 10 |

So, here what we have done in this is my original data set, I add a constant that is 73 every data set is a constant. This sum is equal to 640 that 640, is what I have here the and the mean is 64 which is equal to $59 + 5$. So, adding a constant to every point in the data set is what we have seen so far.

(Refer Slide Time: 20:12)



Statistics for Data Science -1
└Numerical summaries
└Measures of central tendency

Multiplying a constant

$$\bar{y} = \frac{y_1+y_2+ \quad +y_{10}}{10} = \frac{cx_1+cx_2+ \quad +cx_{10}}{10} = \frac{c(x_1+x_2 \quad +x_{10})}{10} = c\bar{x}$$

- Let $y_i = x_i c$ where $c$ is a constant then $\bar{y} = \bar{x}c$
- Example: Recall the marks of students
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66.
  - Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.
  - Then the data becomes
    27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
  - The mean of the new data set is $\frac{236}{10} = 23.6 = 0.4 \times 59$

The next thing is what happens to with the mean, when I multiply each observation of the data set with a constant. Again, let us look at what. So, my $y$ $x_1, x_2, x_n$ is my original data set, I am multiplying it with the constant. Why are we even interested in something of this kind?

For example, I conduct an examination, where the marks are for 100. I conduct an examination for 100; but I want to take only 30% of the grade. So, if the marks when it were for 100 has given me a particular thing, I want to take a particular percentage. So, this $c$ could be a percentage.

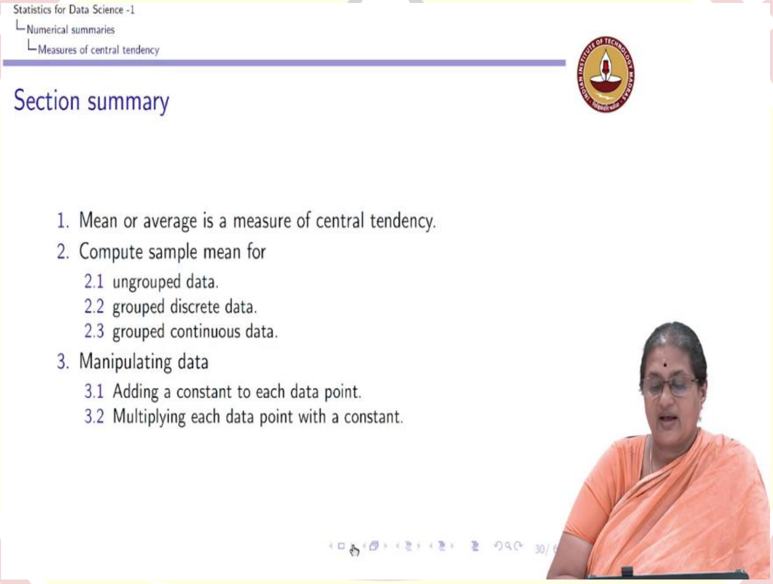
So, if recall I go back to the same example here, I decide to scale down each mark by 40%. In other sense, each mark is multiplied by my $c$ here is 0.4. So, you can see that if I multiply each point $68 \times 0.4$ is 27.2.

So, that is what I have here. I am multiplying each of my data point with 0.4; $68 \times 0.4$, this is $79 \times 0.4$, this is $66 \times 0.4$; I multiply each of my data set with $c$. Now, let us look at what is the sum of that data set is 236 giving me an average of $\frac{236}{10}$ which is equal to 23.6.

Now, you can notice that again, if I go back what is my $\bar{y}$ my $\bar{y}$ is $\frac{y\_1+y\_2+\ my\ y\_10}{10}$, but my $y_1$ is equal to $c$ which times $x_1 + c \times x_2 + c \times x_n$.

This constant is constant for every observation. So, I can remove that constant outside and I can just have $\frac{x_1+x_2+x_3+x_{10}}{10}$. This is $\bar{x}$; hence, $\bar{y}$ is $c \times\ \bar{x}$ ok. So, you can verify my constant here is 0.4; my $\bar{x}$ was 59, so $c \times\ 0.4 \times\ 59$ is what I have is 23.6.

(Refer Slide Time: 23:00)



Statistics for Data Science -1
└Numerical summaries
  └Measures of central tendency

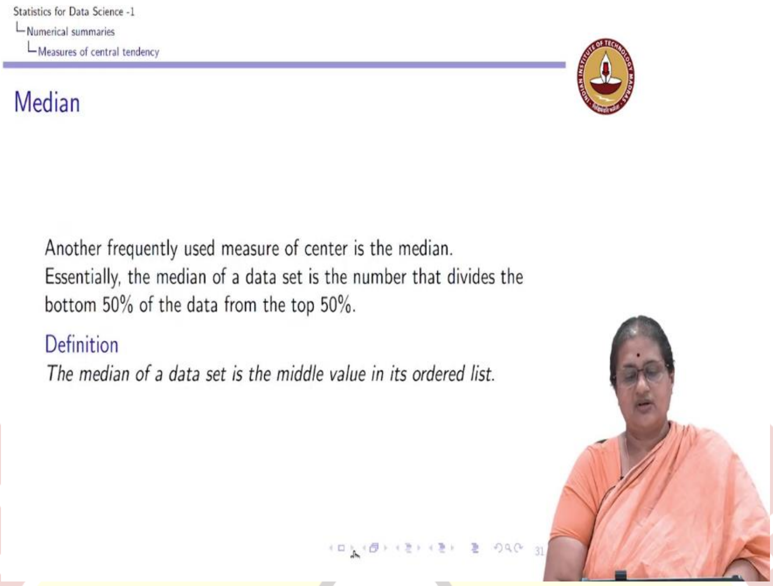
Section summary

1. Mean or average is a measure of central tendency.
2. Compute sample mean for
   2.1 ungrouped data.
   2.2 grouped discrete data.
   2.3 grouped continuous data.
3. Manipulating data
   3.1 Adding a constant to each data point.
   3.2 Multiplying each data point with a constant.

So, what we have learned when we studied about mean is we first looked at what is a definition of a mean, mean is one of the most commonly used summaries of data. We defined what was a sample mean and a population mean. But further, most of the course we are going to only deal with sample mean definition.

We saw how to compute the sample mean for discrete ungrouped data and then, we looked at how to compute the sample mean for grouped data, then we saw what would happen when we manipulate data by adding a constant to each data point and by multiplying each data point with a constant. That is what we have seen so far.

(Refer Slide Time: 23:51)



Statistics for Data Science -1
└Numerical summaries
  └Measures of central tendency

## Median

Another frequently used measure of center is the median. Essentially, the median of a data set is the number that divides the bottom 50% of the data from the top 50%.

Definition

*The median of a data set is the middle value in its ordered list.*

31