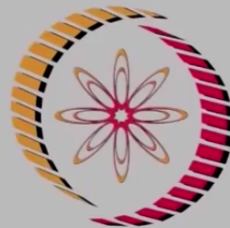




Indian Institute of Technology Kanpur



National Programme on Technology Enhanced Learning (NPTEL)

Course Title
Descriptive Statistics with R Software

Lecture - 17
Central Tendency of data - Mode, Geometric Mean and Harmonic Mean

by
Prof. Shalabh
Department of Mathematics and Statistics
IIT Kanpur

Welcome to the next lecture on the course descriptive statistics with R software. You may recall that in the earlier lecture we had a discussion on computing the quantiles which are the partitioning value and they help us in determining the central tendency of the data. So we will continue on the same topic that what are the different other tools to measure the central tendency of the data and in this lecture we are going to address three more tool that is mode, geometric mean, and harmonic mean.

Mode

Examples:

- A fruit juice shop owner wants to know that which of the fruit juice is more preferred.
- A clothing shop owner wants to know which size of the shirt and/or trouser is highest in demand.

Mode of n observations x_1, x_2, \dots, x_n is the value which occurs the most, compared with all other values.

2

So let us try to start our discussion with mode. So what is here a mode? You might have encountered in your day-to-day life that if you go to a shop of the shirt or shop of the clothings then whenever you want a size of your shirt that is usually available. How it happens suppose you are the shopkeeper and you want to open a shop for the clothings and suppose you want to sell shirts. So which of the size of the shirt is not that much in demand. So you want to know which is the most popular size of the shirts that you should keep in your shop in more quantity. So in this case what you would like to is do is say you will take a sample of the data you will ask people what is the size of the shirt and then you will see that whatever size of the shirt has more frequency that you would try to keep more and the size of the shirt which has a smaller frequency you would try to keep say smaller number of shirts. So this is basically done with the help of mode.

So suppose if fruit juice shop opener wants to know which of the fruit juice is more preferred and simply as I said a clothing shop owner wants to know that size of which of shirt or say trouser is more in demand or say highest in demands and so on. So here in such cases the concept of mode is used. So the mode of say n number of observation is say X_1, X_2, X_n is the value which occurs the most compared with all other values.

Mode

Mode is the value which occurs most frequently in a set of observations.

1, 1, 1, 3, 4, 6, 500
3 1 1 1

Mode is not at all affected by extreme observations.

Distributions having only one mode are called unimodal and the one with two modes is called bimodal.



So essentially the mode is the value which occurs most frequently in the set of observation. How this frequently word is coming into picture that is coming because of the frequency distribution or the frequency of the values. So the definition of the mode will be interconnected with the frequency of those observation or the frequency distribution of those observations. One advantage of mode is that mode is not at all affected by the extreme observations. For example, if I say I have data set here 1, 1, 1, and say here 3, 4, and 6 so you can see here that number 1 is occurring here three times, 3 is occurring one time, 4 is occurring one time and 6 is occurring one time. So the maximum number of value which is occurring here this is here 1. So the mode here is going to be 1. Now in case if I try to add here a value 500 mode is not going to be changed because that is also appearing further one time only. So that is the basic idea that mode is not at all affected by the extreme observations and in case if you try to plot the frequency curve and suppose I have these two types of frequency curve. One is going to give like this and say another is going to give like this. So you can say here that here is only one mode but here you can see here although the more here is only one but still there are two peaks so we call associate contribution as a bimodal and in first case they are called as unimodal. So all the distribution having only one mode they are called unimodal and all those distribution which have two modes it's called as bimodal.

So now we try to define the mode for two cases when we have group data and when we have ungroup data. Here I would try to inform you that in the R software and in the base package of R software there is no direct command to find out the mode. Well there is a command mode, m-o-d-e but be careful that mode is not used to find out the mode that we are trying to find as a measure of central tendency. That command is used to describe the behavior of the data that means that data is where the numeric or not something like this. So please be careful. Although I will try to show you here that by writing some special functions or say some special commands we can find out the mode. But you cannot use the function m-o-d-e or the command m-o-d-e which is built-in inside the R software. So be careful.

Mode for Ungrouped Data

For discrete variables, the mode of a variable is the value of the variable having the highest frequency in a unimodal distribution.



So first we try to discuss the mode for ungrouped data. So for the ungroup data or for the discrete variables mode is very simple. The mode of a variable is the value of the variable which has got the highest frequency and obviously this is true in the case of unimodal distribution. So what you have to do here you will have here say X_1 data with frequency f_1 . X_2 data with frequency f_2 and so on X_n data with frequency f_n and you simply have to first choose that whichever is the maximum value. Suppose this maximum value occurs as say f_m then corresponding to f_m whatever is the value of here X_m that is going to be the mode. So obviously this is true when we have a unimodal distribution where there is only one mode.

Mode for Ungrouped Data

R command

Step 1: Create a table of given data vector or matrix `data` say `modetab` as follows:

```
modetab = table(as.vector(data))
```

The first row of `modetab` is a sorted list of all unique values in `data`.

Step 2: Following returns the names of the values having the highest count in second row of `modetab` which is the mode.

```
names(modetab)[modetab == max(modetab)]
```

< > == logical operator, logical equal

So in order to find out this mode in the R software we will go like this. If you try to understand I am simply going to write here two steps and they are simply trying to copy the same thing what I had just told you that you try to create a frequency distribution, try to choose the maximum value of the frequency and corresponding to the maximum value of the frequency try to choose the value of X and that is going to give you the value of your mode. So when we try to compute this mode in R software I am giving you here two steps. Step one and step two. The step one is very simple whatever is your data try to create a table of that data vector or that can be a matrix also. How that will be useful? I will try to show you. So whatever is your data either in the form of a vector or a matrix try to convert it into a table. So I will try to store my data inside a variable named Data and whatever is the outcome of the table of this data yes I am going to indicate as modetab. Yeah that is the short form of mode of table data.

So in order to use this thing I will use here a command table that we had used earlier and inside the arguments I will try to write here as.vector as.vector that means you try to create or you try to consider that the data is vector and this data which has to be considered as vector this is given here inside the variable named data. So I will show you later on but here I can tell you that whatever is the outcome of this command the first row of this command will be a sorted list of all the unique values in the data vector data. Now after this you have to operate here over one more command. This command I am writing here. This is here names then inside the arguments you have to write the data that you have obtained in the first step. This data is called as modetab then inside the square brackets, remember these are the square brackets here. Inside this bracket you have to write modetab or the data what you have obtained in there step one is double equal to, what is this two equality sign, this is a sign of a logical operator for equality. So that is a sign for logical equal sign for example we have less than sign, greater than sign, and equality sign but equality sign means equal to but the logical equal sign is denoted by two equality signs. And then I am trying to find out the maximum value of the data which is inside the modetab.

Here I would just like to inform you that here I have used the commands here names of modetab and this function which is generally taught when you try to learn the R software logical operators and how to find or extract the names from R data and so on but here I would not going into those details but I would simply request you please try to use this command. So now suppose I simply try to take an example here and try to show you that how these things are happening.

Mode for Ungrouped Data
R command

```
> data = c(10, 10, 10, 10, 2, 2, 3, 4, 5, 6)
```

Create a table of given data vector data

```
> modetab = table(as.vector(data))
```

```
> modetab
```

2	3	4	5	6	10
2	1	1	1	1	4

sorted list of all unique values in data

names of values having highest count

freq → 4

```
> names(modetab)[modetab == max(modetab)]
```

```
[1] "10"
```

So I am simply trying to pick up data which is here like this inside the data vector and from there I am trying to create the table of this data using the same command and here is the outcome. You can see here. So you can see here that here are in the data vector of the values are 10, 10, 10, 10, 2, 2, 3, 4, 5, 6. So essentially when we try to create the frequency table so it is trying to show you in the first row these are the sorted values of all the unique values in data. You can see here that the unique values here are 2, then here 3, then here 4, then here 5, then here 6 and this value is repeated again and again four times that is 10 and now in the second row this is trying to count that how many times a value is occurring. For example if you try to see here 2 is occurring two times. One, two. So this is here two. 3 is occurring here only one time so this is here one. 4 is occurring here only one time. So this is here one. 5 is occurring here say one time. So this is here one. Now and then 6 is occurring here say one time this is here one. And now 10 is occurring here 1, 2, 3 and here 4 time so this is here 4 and this is here the 10. So what essentially I need to do in order to find out the mode I simply have to extract the values from this frequency first.

Mode for Ungrouped Data

R command

```
R Console
> data = c(10,10,10,10,2,2,3,4,5,6)
>
> modetab = table(as.vector(data))
> modetab

 2  3  4  5  6 10
2  1  1  1  1  4
>
> names(modetab)[modetab == max(modetab)]
[1] "10"
```

So the frequency, maximum frequency is at 4 this here and then in this second step I have to find out the value corresponding to this maximum frequency which is here 10. So this command here which is here at the second steps this tries to extract this value 10 from the first row and this gives me here 10. So this is how you can compute the mode but definitely I would like to address here this is not the only way out. You can define it at an own way is using your own logic also. And this is here the screenshot and I would request you that you please try to execute it on your data also. So for example I can show you it on the R console but you please unless and until you do it with your own hand on the R console you may not really understand it. So I have got our data you can see here and with this data I am trying to get the value here modetab so you can see here this is here modetab and now after this modetab I'm using the command in the step two to get the value of the mode. This is here ten.

Mode for Ungrouped Data

R command

```
> data = matrix(nrow= 3, ncol=3, data=c(1,2,2,  
3,3,4,5,6,6))
```

```
> data
```

	[,1]	[,2]	[,3]
[1,]	1	3	5
[2,]	2	3	6
[3,]	2	4	6

mode 2 mode 3 mode 6

mode

Now I will try to show you another aspect. As I told you that this command which I just introduced that can be used on a data vector or a data matrix. So suppose I have a data in the form of a matrix and we have learned how to write the data inside the matrix. So there is a three by three matrix and the data values are given inside this data vector. The values are 1, 2, 2, double 3, 4, 5, double 6 so this matrix will look like this. When I say I would like to find out the mode of this matrix essentially I am trying to say that I need the mode of first column, second column, and third column. So in the first column you can see here the value 2 is occurring two times so here the mode should be equal to 2. And the second column the value 3 is occurring two time. So here the mode should be equal to 3 and in the third column the value 6 is occurring two times so here the mode should be equal to 6.

Mode for Ungrouped Data

R command

Create a table of given data matrix `data`

```
> modetab = table(as.vector(data))
```

```
> modetab
```

```
1 2 3 4 5 6  
1 2 2 1 1 2
```

```
> names(modetab) [modetab == max(modetab) ]
```

```
[1] "2" "3" "6"
```

So this is what I mean when I try to repeat the same command on this data matrix. So I try to do the same thing I have to simply copy and paste the same thing so you can see here I'm using the same command but now it is giving me here this type of value and then I'm trying to use here the command in the second step and it is giving me an outcome here 2, 3, 6. So you can see here this 2, 3, 6 is the same what you had obtained manually here 2, 3 and here 6.

So my idea was simply to inform you that this data can be a data vector or a data matrix and here is the screenshot of the same operation that we done.

Mode for Grouped Data

For continuous variable, the mode is the value of the variable with the highest frequency density corresponding to the ideal distribution which would be obtained if the total frequency were increased indefinitely and if, at the same time, the width of the class intervals were decreased indefinitely.



11

And now we come on the aspect that how to compute the mode for the grouped data or the data when is on continuous-wave variable. So in the case of continuous variable the mode of the data is the value of the variable with the highest frequency density corresponding to the ideal distribution. What is an ideal distribution? Which would be obtained if the total frequency were increased indefinitely that will be they are becoming very very large and if at the same time the width of that class intervals were decrease indefinitely. Now you may recall that we had a discussion on histogram and frequency curve. What we had seen that frequency curve or the frequency density or density plots they are more useful when you have large number of data and in that case the bins of the histograms are reduced and they are made as small as possible and the number of data points are made as large as possible. So this is the same thing which is trying to say so under that thing in case if you are trying to make here a frequency curve like this the bins are going to be very very small and so on and then you will get to here at this type of frequency curve. So this is the highest value of the frequency around which you will get the value of the mode.

Mode for Grouped Data

<u>Class intervals</u>	<u>Mid point (m_i)</u>	<u>Absolute frequency (f_i)</u>
$e_1 - e_2$	$m_1 = (e_1 + e_2)/2$	f_1
$e_2 - e_3$	$m_2 = (e_2 + e_3)/2$	f_2
...
$e_{K-1} - e_K$	$m_K = (e_{K-1} + e_K)/2$	f_K

$\sum_{i=1}^K f_i = n$

max $\rightarrow f_m$
 \downarrow
 m

Modal class: Class corresponding to the maximum frequency.

Now in order to compute the mode for this grouped data the first step is to create a frequency table and this frequency table we just need three things. One is the class intervals. Second is the midpoint of the class intervals. And the third is here the absolute frequency. One thing what you have to notice here that in this case I am trying to use the symbol f to denote that absolute frequency and earlier in some of the lectures I have used the symbol f to denote the a relative frequency but I am just trying to keep the standard symbols so that you don't face any problem once you try to read from the books.

So here if you try to see I have created different class intervals. Now I am simply trying to find out the midpoint of this class interval say m_1, m_2, m_k and they are obtained simply by finding out the value of lower limit plus upper limit divided by 2 and corresponding to the first class I have frequency f_1 . Corresponding to second class I have frequency f_2 and corresponding the k th class I have frequency f_k . Now what we have to do? We simply have to find out the maximum value among these frequencies and whatever is the maximum value say here f_m I have to identify wherever this f_m is lying corresponding to which I have to find out the value of here m and that will give me the based on that I will try to compute the value of the mode. So the class where this maximum frequency is occurring this is class modal class in order to compute the mode for the group data the expression is given by like this.

Mode for Grouped Data

$$\bar{x}_{mo} = e_l + \frac{f_0 + f_{-1}}{(f_0 - f_1) + (f_0 - f_{-1})} d_l$$

e_l : lower limit of modal class

d_l : class width

f_0 : frequency of modal class

f_{-1} : frequency of the class just before the modal class

f_1 : frequency of the class just after the modal class

Class 1 $\rightarrow f_{-1}$
Am $\rightarrow f_0$
Class 2 $\rightarrow f_1$

Well this is based on certain computation certain derivation but I am not going into that detail.

So here this value here e_l is the lower limit of the modal class and this d_l it is the class width and f_0 here is the frequency of the modal class and by this f and inside the subscript it is minus 1 this is denoting the frequency of the class just below the modal class and this frequency here f_1 this is just indicating the frequency of the class just after the modal class. So for example if I have here modal class here see here Am and then I have here 1 say here class, say here 1 and this here is class 2 then this is going to the frequency of this Am is going to be of, this is going to be $f-1$ and after this this is going to be f_1 .

Mode for Grouped Data

Example

The time (in minutes) taken by a customer to arrive in a shop in a month on different days are recorded as follows:

Day	1	2	3	4	5	6	7	8	9	10	
No. of minutes	30	31	30	30	29	29	29	29	29	28	
Day	11	12	13	14	15	16	17	18	19	20	
No. of minutes	28	28	27	27	27	26	26	26	26	25	
Day	21	22	23	24	25	26	27	28	29	30	31
No. of minutes	25	25	25	25	25	24	24	23	22	21	21

So not based on that we will try to compute the mode. Now we consider an example and I will try to show you that how you are going to compute this data the mode based on this data. So this is again here the same example that we considered earlier that the time taken by a customer to arrive in our shop in inside a mall on different days is recorded and there are 31 days so there are 31 values here on the number of minutes that the customer takes.

Mode for Grouped Data

Example: Considering the data as grouped data, we can present the data as

Class intervals	Mid point (x_i)	Absolute frequency (f_i)
15 – 20	$\frac{15+20}{2}$ 17.5	$f_1 = 0$
20 – 25	22.5	$f_2 = 12$
25 – 30	27.5	$f_3 = 18$
30 – 35	32.5	$f_4 = 1$
35 – 40	37.5	$f_5 = 0$

Modal class: Class corresponding to the maximum frequency.
 $l = 3 : 25 - 30$

Now we try to convert or we try to prepare the frequency table. So you can see here I have made here five classes, five class intervals and I have computed their midpoint for example in this case the midpoint is computed as 15 plus 20 divided by 2 is equal to 17.5 and a frequency is here zero and similarly other midpoints have been calculated and their corresponding frequencies have been calculated. Now in this case out of this frequency f_3 equal to 18 this is the maximum frequency which is occurring. So the modal class is going to be the class corresponding to which we have the maximum frequency. So here this 25 to 30 is going to be the modal class. So I can take here that l is equal to 3 which is 25 to 30 this interval this is the modal class. Now based on that I will try to see here for example here you can see the frequency of the modal class is something like what we had denoted as f_0 . This is 18 and the frequency just for the modal class which is here f_{-1} as per our notation is 12 and the frequency just after the modal class which is here denoted as f_1 this is equal to here 1. So substituting these values we will try to find out the value of the mode from the given expression.

Mode for Grouped Data

Example:

$e_l = 25$: lower limit of modal class

$d_l = 5$: class width

$f_0 = 18$: frequency of modal class

$f_{-1} = 12$ frequency of the class just before the modal class

$f_1 = 1$: frequency of the class just after the modal class

$$\begin{aligned}\bar{x}_{mo} &= e_l + \frac{f_0 + f_{-1}}{(f_0 - f_1) + (f_0 - f_{-1})} d_l \\ &= 25 + \frac{18 + 12}{(18 - 1) + (18 - 12)} \times 5 \approx 31.52\end{aligned}$$

16

So you can see here that e_l is the lower limit of the modal class which is 25. The width of the class is 5 and these are the frequencies modal class frequencies f_0 . The frequency of the class just before the modal class as 12 and frequency of the class just after the modal class as 1 and if you try to substitute all these values over here you get here the value 31.52. Now I would like to address here one thing first. Inside R there is no built-in function to compute the mode of the grouped data what we have just obtained through formula. Well you can write a small function or a small program to compute it but definitely I am not going to consider with that idea over here. And now I would try to address two more tools geometric mean and harmonic mean which are also two different measures to find out the central tendency of the data. So first I try to address geometric mean.

Geometric Mean

Geometric mean is useful in calculating the average value of ratio or rate of interest etc.

Not applicable of any of the observation is zero.

17

Geometric mean is useful in calculating the average value of ratios or rates of interest in say banking and in finance sector and so on and this geometric mean is not really applicable when any of the observation is 0. Why?

Geometric Mean

x_1, x_2, \dots, x_n observations which are all positive.

The geometric mean for

- Ungrouped or discrete data is

$$\bar{x}_G = (x_1 \times x_2 \times \dots \times x_n)^{\frac{1}{n}}$$

$x_i \rightarrow \text{freq} = f_i$

- Grouped or continuous data with frequency distribution is

$$\bar{x}_G = (x_1^{f_1} \times x_2^{f_2} \times \dots \times x_n^{f_n})^{\frac{1}{N}} \quad \text{where } N = \sum_{i=1}^n f_i$$

where x_1, x_2, \dots, x_n occur with frequencies f_1, f_2, \dots, f_n respectively.

18

This is going to be clear from the definition of the geometric mean. So one condition in geometric mean is that all the observations should be positive. So let X_1, X_2, X_n be the n observations which are all greater than 0. Now these observations can be on a discrete variable or an ungrouped variable data set or they can be data on continuous variable and may create a grouped data. So when these observations are ungrouped then in this case the geometric mean is defined by here \bar{X}_G something like this so what we are trying to do here we are simply trying to take all the observation X_1, X_2, X_n we are trying to multiply it and then I am trying to find out the n th root by taking the power to be here 1 upon n . And similarly in case if you have the grouped the data in which every X_i has frequency see her f_i then in this case the geometric mean is defined by X_1 raised to the power of f_1, X_2 raised power of f_2, X_3 raised power of f_n and all these values have to be multiplied. So this is the product of $X_i^{f_i}$ raised to the power of f_i and here the power here is 1 upon capital N . N is the sum of all the frequencies. So now in case if you want to find out the geometric mean in the R software once again there is no direct command but writing down this command is very simple. If you remember in the initial lectures we have discussed different types of built-in functions and using those built-in functions we can create the command for the geometric mean. So if you see here I have created this command here but there are different ways to construct it.

Geometric Mean for Ungrouped Data
R Command

x : Data vector

Geometric mean for discrete data

prod(x)^(1/length(x))

(length(x) is equal to the number of elements in x)

So suppose if I try to denote the data vector here as say X then if you try to see what are we going to do in the case of ungrouped data first I am trying to find out the product of all the observation and then I am trying to take the power 1 upon here n . So this product can be stored or can be defined by product of X where X is containing all this data and then for this power I am using here hat and then this here n , n here is the number of observations in X . So this can be

determined by lengths of X. So this power can be written as hat and inside the bracket 1 upon length of X.

Geometric Mean for Grouped Data
R Command

x: Data vector $c(x_1, x_2, \dots, x_n)$

f: Frequency vector $c(f_1, f_2, \dots, f_n)$

where x_1, x_2, \dots, x_n occur with frequencies f_1, f_2, \dots, f_n respectively.

Geometric mean for continuous data

$\text{prod}(x^f)^{(1/\text{sum}(f))}$

$(x_1^{f_1} \times x_2^{f_2} \times \dots \times x_n^{f_n})^{1/\sum f_i}$

(sum(f) is equal to the sum of elements in f)

20

So this is pretty simple and similarly in case if you have a group data where you are trying to say that the data vector has something like values X_1, X_2, X_n with frequencies f_1, f_2, f_n such that the X_1 has frequency f_1 , X_2 has frequency f_2 and X_n has frequency f_n . So I can write the data vector and the frequency inside a vector so that data is written inside the data vector X and all the frequencies have been written inside the vector here f and finding out these frequencies is not difficult just by creating a frequency table and extracting the frequencies as we have done it earlier one can obtain it.

So now whatever is the product something like X_1 raised to the power of f_1 multiplied by X_2 raised to the power of f_2 and so on multiplied by X_n raised to the power of f_n this can be written by product of X raised to the power here f. And then it is here 1 upon N. N is going to be the sum of all f_i . So sum of all f_i this can be determined by the function sum of f and so writing it here like this you will get the value of the geometric mean in this case. So you can see here that just by using the built-in functions in R you can find out the geometric mean.

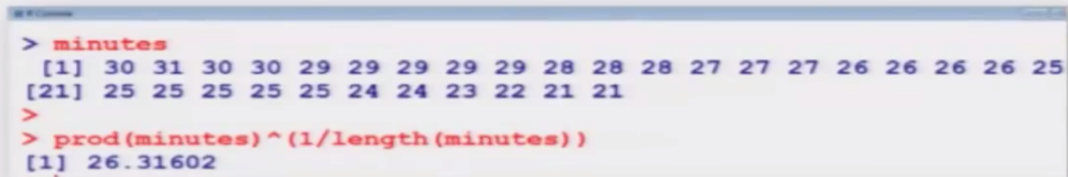
Geometric Mean for Ungrouped Data

Example: Considering it as ungrouped data

```
minutes = c(30,31,30,30,29,29,29,29,29,28,28,  
28,27,27,27,26,26,26,26,25,25,25,25,25,25,24,2  
4,23,22,21,21)
```

Geometric mean for discrete data

```
> prod(minutes)^(1/length(minutes))  
[1] 26.31602
```



```
> minutes  
[1] 30 31 30 30 29 29 29 29 29 28 28 28 27 27 27 26 26 26 26 25  
[21] 25 25 25 25 25 24 24 23 22 21 21  
>  
> prod(minutes)^(1/length(minutes))  
[1] 26.31602
```

21

So now in case if I try to take an example the same example in which I have considered the data on minutes so this data is here contained here like this and if you want to find out the geometric mean considering this data as a discrete data you simply have to use this expression and you can see here this is the value of the geometric mean and this is here the screenshot and considering this data as a group data we can find out this frequency table that we already have discussed and based on that I have this data here that is the same data.

Geometric Mean for Grouped Data

R command

Example

Frequency distribution

```
> breaks = seq(15, 40, by=5) # sequence at  
interval of 5 integers  
  
> breaks  
[1] 15 20 25 30 35 40  
  
> minutes.cut = cut(minutes, breaks, right=FALSE)  
  
> minutes.cut  
[1] [30,35) [30,35) [30,35) [30,35) [25,30) [25,30) [25,30) [25,30)  
[9] [25,30) [25,30) [25,30) [25,30) [25,30) [25,30) [25,30) [25,30)  
[17] [25,30) [25,30) [25,30) [25,30) [25,30) [25,30) [25,30) [25,30)  
[25] [25,30) [20,25) [20,25) [20,25) [20,25) [20,25) [20,25)  
Levels: [15,20) [20,25) [25,30) [30,35) [35,40)
```

And now first I need to extract the frequencies. How it can be done? You may recall that in the earlier lecture in the case of median I had shown you that how to find out the or how to extract the frequencies from a group data. So I am not going to explain it here again but I will simply be using the same command that I had used in the case of median and if you have forgotten I would request you please try to go to the lecture on the median and try to see that how these values like as breaks, cut commands were used to extract that data. So now using the earlier explained method I will try to construct here a data vector breaks which is a sequence of 15 to 40 at an interval of 5 because our class intervals are starting from 15 and going up to 40 and they are of the width 5. So this data comes out to be as 15, 20, 25, 30, 35, 40, and then I have to operate the command here cut on the given data vector minutes using here the breaks and right intervals are going to be open so this is right equal to false and once you try to do it I will get here this type of data. So you can see here these are the class intervals that we have obtained. Now I have to find out the frequencies of this minutes.cut data.

Geometric Mean for Grouped Data

R command Example

Frequency distribution

```
> table(minutes.cut)
```

```
minutes.cut
```

```
[15,20) [20,25) [25,30) [30,35) [35,40) → class interval  
0 6 21 4 0 → freq
```

Extract frequencies from frequency table using command

```
as.numeric(frequency table data)
```

```
> f = as.numeric(table(minutes.cut))
```

```
> f
```

```
[1] 0 6 21 4 0
```

25

So in order to find out the frequencies I will simply operate here the command table and inside the arguments the data that is minutes.cut it will give me the frequency table. So these are the – the first row is indicating the class interval and the second row is denoting the frequencies.

Geometric Mean for Grouped Data

R command

Example

```
> x = c(17.5,22.5,27.5,32.5,37.5) # Mid values
```

```
> f = as.numeric(table(minutes.cut))
```

```
> f
```

```
[1] 0 6 21 4 0
```

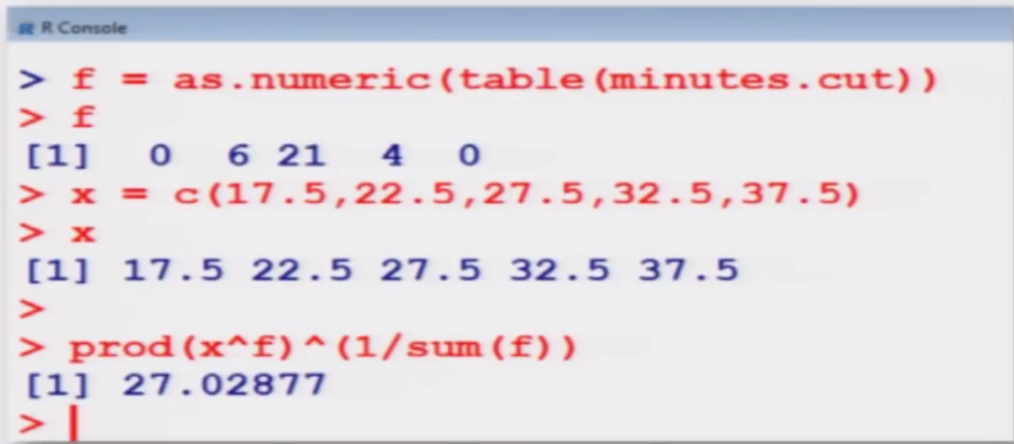
Geometric mean for continuous data

```
> prod(x^f)^(1/sum(f))
```

```
[1] 27.02877
```

26

Now how to extract this frequency vector for that we have used the command as numeric and I try to store this value here in this f so as numeric on the data provided by table minutes.cut and this comes out to be like this here. So now I have obtained here the vector f and now I have to collect all the midpoints and I have to create a vector here X. So I am trying to collect all the midpoint 17.5, 22.5, and so on and now I have here two vectors X and here f and based on that I can use this function which we have just evolved to compute the geometric mean. And here if you try to see I have given here the screenshot but I would request that you please simply try to copy these commands, paste it into your R console and try to see whether you are getting the same outcome or not. And this is the same outcome which you will be getting here.



Mode for Grouped Data
R command of mode
Example

```
R Console
> f = as.numeric(table(minutes.cut))
> f
[1] 0 6 21 4 0
> x = c(17.5,22.5,27.5,32.5,37.5)
> x
[1] 17.5 22.5 27.5 32.5 37.5
>
> prod(x^f)^(1/sum(f))
[1] 27.02877
> |
```

28

Now after this I will come to the last topic on the measure of central tendency that is harmonic mean. So harmonic mean is also defined for group data and say ungrouped data. For our discrete data the harmonic mean is defined here as say here \bar{X}_H which is equal to $\frac{1}{\frac{1}{N} \sum \frac{1}{X_i}}$. So doesn't it look like if I am trying to find out the mean of the inverse of the observations and then I am trying to take it here once again the inverse and the same definition is for the continuous data for the group data and the expressions for finding out the harmonic mean is like here this. But here you have to see here that X_i has frequency f_i same terminology that we have used in the case of geometric mean in the grouped data case.

Harmonic mean

Observations: x_1, x_2, \dots, x_n

For discrete data

$$\bar{x}_H = \frac{1}{\frac{1}{n} \sum_{i=1}^n \left(\frac{1}{x_i} \right)} \quad \left. \vphantom{\bar{x}_H} \right\} \rightarrow \frac{1}{\text{mean}\left(\frac{1}{x_i}\right)}$$

For continuous data having frequency distribution

where $N = \sum_{i=1}^n f_i$

$$\bar{x}_H = \frac{1}{\frac{1}{N} \sum_{i=1}^n \left(\frac{f_i}{x_i} \right)} \quad \left. \vphantom{\bar{x}_H} \right\} \rightarrow x_i \text{ has freq } f_i$$

So now in case if you want to compute the harmonic mean in the R software once again I would like to inform you that there is no built-in command inside the base package of R but writing down a small command to compute the harmonic mean is not difficult. Just by using the built-in function and by looking at the structure how the mean has been defined one can easily do it.

Harmonic mean for discrete data

R Command

x : Data vector

Harmonic mean for discrete data

$1/\text{mean}(1/x)$

$$\frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}} = \frac{1}{\frac{1}{n} \sum_{i=1}^n y_i} = \frac{1}{\text{mean}(y)}$$
$$= \frac{1}{\text{mean}\left(\frac{1}{x}\right)}$$

So if you try to see here the command which I am writing here is for this quantity $1/\text{mean}(1/x)$. So you can see here if I try to denote here $1/x_i$ to be here something like here y_i then this quantity becomes $1/\text{mean}(y)$ and which is nothing but $1/\text{mean}(1/x)$. So this can be written here simply here $1/\text{mean}(1/x)$ see here $1/x$ vector in you R software. So that is the same thing which I am writing here that if your X is a data vector then the harmonic mean in case of discrete data is defined as $1/\text{mean}(1/x)$.

Harmonic mean for continuous data

R Command

x : Data vector $c(x_1, x_2, \dots, x_n)$

f : Frequency vector $c(f_1, f_2, \dots, f_n)$

where x_1, x_2, \dots, x_n occur with frequencies f_1, f_2, \dots, f_n

respectively.

Harmonic mean for continuous data

$1/\text{mean}(f/x)$

$$\sum \left(\frac{f_i}{x_i} \right) \rightarrow (f/x)$$

31

And similarly in case if you have a continuous data and the group format we are X_1, X_2, X_n are the values which are occurring with the frequencies f_1, f_2, \dots respectively then the harmonic mean once again can be defined by the 1 upon mean of f upon X because if you try to see you are simply trying to compute the average of this $f_i X_i$ so for that f_i of X_i this is written here to say f divided by X in the R symbol and then I'm simply trying to find out here the mean of this and then I am trying to invert it.

Harmonic mean

Example

```
minutes = c(30,31,30,30,29,29,29,29,29,28,28,
28,27,27,27,26,26,26,26,25,25,25,25,25,25,24,2
4,23,22,21,21)
```

Harmonic mean for discrete data

```
> 1/mean(1/minutes)
[1] 26.17633
```

32

So now if I try to take the same example that we considered earlier of the minutes so this is here the data which I am stored in the variable minutes and if you simply try to execute this data upon this command to compute the harmonic mean that we just discuss you will get this value over here.

Harmonic mean

Example

```
> x = c(17.5,22.5,27.5,32.5,37.5) # Mid values
> f = as.numeric(table(minutes.cut))
> f
[1] 0 6 21 4 0
```

mid points

Harmonic mean for continuous data

```
> 1/mean(f/x)
[1] 4.335085
```

f
x =

33

This is not difficult at all now and similarly in case if you want to consider this data as a continuous data so as we had obtained the frequencies in the geometric mean case up to that point you have to copy that the same thing and finally we were getting the frequencies like as here. So now you have got here the f vector and you have got here the X vector. X is here this – this is the midpoints. So this is what you have to keep in mind that in this case X is are the midpoints and if you try to just execute this command `l upon mean of F upon x` then you will get this value here. And these things are not very difficult to obtain. You can see here this is the screenshot of the same thing.

Now I would like to stop here in this lecture and I will also like to stop on the topics of measures of central tendencies. So we have discussed in this chapter different types of tool, arithmetic mean, geometric mean, harmonic mean, median, mode, for the group data, for the ungroup data and as far as possible wherever available I have explained you how to compute it in the R software. So once again I would request you that you take different types of data set and on the same data sets you can compute each and everything, this quantiles, mean, median, mode, harmonic mean, geometric mean, try to convert the same data into group data as well as ungroup data try to operate the things and try to see how much difference you are getting and try to think why this difference is coming that you will get from the theory of statistics. So you practice it and we will see you in the next lecture. Till then good bye.