

Lecture 08

Absolute Frequency, Relative frequency and frequency Distribution

Welcome to the, next lecture on the course, descriptive statistics with our software. you may recall that, in the earlier lectures, we had discussed, different aspects, related to statistics. And we have understood that

whenever, we want to do any statistical analysis. How are we going to start? And how we are going to control, the process of obtaining the, observations, that different types of associated variables, discrete continuous etcetera? So, now I assume that, we have collected a sample. And as I said earlier, I will always be assuming that, my sample is representative. Which means, that all the salient features, which are present in the population, they are also present in my sample? So now, we are at a place, where we have collected the observations. And now, we want to move further. So, first of all, whenever the data comes to our hand, as I told you, that there are two options, we can start, with one is graphical tool and another is analytical tool. So, first of all I would try to take the first analytical tool, which will give us an idea, that how are we going to combine, the data present the data and how we can do it, which will give us, some information, that is contained inside the data. And based on that, we will try to take, further decisions, that what type of graphical tools and tools, for analysis of the data, can be used. So, we start this lecture and first we are going to address, that once the data comes, why it needs to be classified, you can always assume, inside your mind, that whenever you conduct an experiment, you will get the data. Now, I am also assuming, that the data is collected on the relevant variable, which you want to study, for example, if you want to stay the height, that then the data is on the height, if you want to either weight, becomes, data is on the weight. So, now you have collected the data, this data can be 20 observations, 100 observations, 200 observations, 2 Millions observation, 2 billions observations. So, in case, if all the values are just before you, can you really, get an idea of, what is the information had an inside it? It is very difficult. Because, as I said, data cannot speak, data cannot come outside, of your computer or outside your experiment, to tell you that. Okay? I have this piece of information, this is only you, who has to use appropriate tool to get it out, to take it out. So, first thing what we try to do? We try to rearrange, the data in some required format,

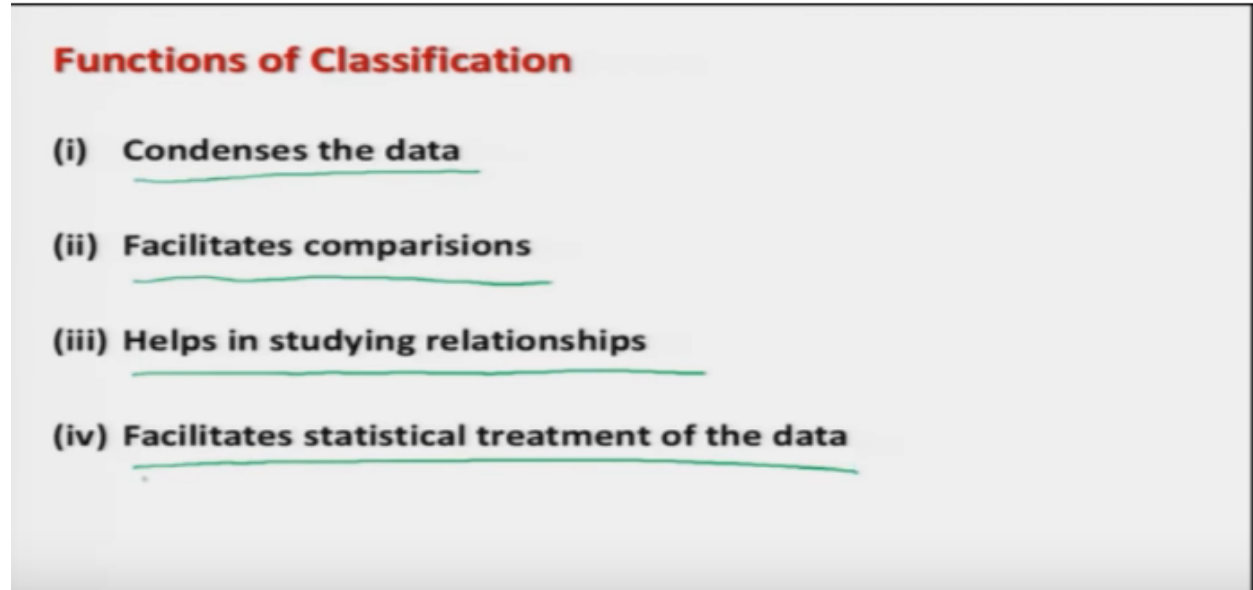
Refer slide time: (3:32)

Classification of Data

Process of arranging the data into groups or classes according to resemblance and similarities.

And for that, we would like to classify, the data into, different groups and from different aspects. For example, I can make the groups, of those observations, which are similar. Or which are also dissimilar, all those units, which looks similar to each other, they can be put into one group, similarly all those units, which looks, similar they can be put in another group. And then, based on that, we can extract the information, different types of information, through those groups. So, this is what we are going to, now study, so the classification of the data consists of a very simple thing, that this is a simple process of arranging, the data into groups or classes, according to resemblance and similarities. Okay?

Refer slide time: (4:27)



Now, what are the functions of, for this classification, why do we make this classification? The biggest advantage is that, this will condense the data, you can assume that on, one of the walls, different numbers are written, continuously. And if you try to look at those numbers, there are thousands and ten thousand, 1 million numbers, you can't, get any information out of that. So, you need to condense the data, so one of the important objective, of classification is this, that we would like to, condense the data. Contain the data in class away, from where we can, we draw some relevant information. And whenever you are trying to make a stat scale experiment, generally your objective is to compare something, for example, if there is a new medicine, which claims that it, can control the body temperature, for say, for say 12 hours , then you would like to compare it, with the ,with the earlier existing medicine. That whether this improvement is, happening or not? So, this condensation of the data or the classification of the data has to be done, in such a way, which can help us, in comparing different types of things, different types of aspect, different types of quantities, different types of natures. And in many many situations, usually we are interested in studying the relationship, for example, modeling, aesthetical modeling is a very popular word. Which is nothing but, a sort of relationship? We want to find out the relationship between, input and output variables. So, the models cannot be obtained in a single shot, the models are obtained on the basis obstetrical data and this is the starting point, means about, descriptive statistics ,from there we try to gather, the data information in a small pieces and then, we try to, combine them together, in getting a model. So, we would like to condense the data in such a way, which helps us, in studying different types of relationships. And the data has to be condensed in such a way or the data has to be grouped in such a way that is compatible with our sterile tool also. This is that thing, what we always have to keep in mind. Usually people do what? That first they will collect the data and then they will try to choose the statistical tool, what I always suggest is that, you first try to fix your objective and then try to see, what type of statistical tool, can be used to fulfill or to give an answer, of that objective and whatever is the requirement of that tool, try to collect your data, according to that and this will, help us. So, another important function of classification is that, this should facilitate the statistical treatment of the data.

Refer slide time: (7:29)

Absolute and relative frequencies

Suppose there are 10 persons participated in a test and there results were declared in two categories as Pass (P) and Fail (F).

P, F, P, F, F, P, P, F, P, P

Use a_1 and a_2 to refer to Pass and Fail categories.

There are 6 persons who passed, denoted as $n_1 = 6$.

There are 4 persons who failed, denoted as $n_2 = 4$.

The number of observations in a particular category is called the

absolute frequency

12:16 / 46:23

So now, moving further, first till, then let me introduce, the basic definition. This is absolute and relative frequencies. One thing I would like to make it clear here, that in, order to teach in this course, I have two option, that first I try to take the theory, formulas etc. And then I try to, pick an example. But, rather I would prefer in most of the situation, that I should start, at an example and then, I try to develop the theory so that, you can make a one-to-one correspondence, between the theory and the, and data definitions. That will help you, in applying or choosing, the tools in, R software. Okay? So, now let me take a simple example, suppose there are, ten persons, who participated in a test and their results were declared, their results were declared in two categories, either they passed or they failed. so the candidate, who has passed, he or she has been assigned, the letter capital P and the candidate, who got failed he or she has been assigned L greater capital F. so, now you can see here, that this is here ,the data of 10%, who either got pass or fail and their outcome is recorded here as say p, F,P, f, F,P, p, f, P,P. well, I'm trying to take care of very small Decatur said, which you can see from your eyes and what about mathematical manipulations, I am doing, that you should be able to see from your, own eyes. But, you can always think that, this data can be very very large, there can be ten thousand candidates, there can be 1 million candidate, there can be 10 million candidates. So, now how to combine this data? How to condense this data? So, we are going to use the, concept of absolute frequencies and relative frequencies, to condense the data. And then later on we will try to put them, in some proper format for example, in the form of a table, to get more clear information. So, now I would try to, did we note here, there are two categories, there are two categories. One is here pass and say, another is air fail. So, I can now in general, represent these categories as say here, a 1 and here a 2. So, this a 1 category will represent the candidates who have passed? And a 2 category will represent those candidate who have failed? So, now I have, introduced here word category. So now you can see, category contains all the observation, which are similar to each other, for example, the category of candidate who pass this will contain all the candidate, who have passed? The

category fail, contains all the candidates, who got failed in the test. Right? So these are called, 'Categories'. So, I can see here, that there are, some number of candidates who passed and some number of candidate who failed. So, let me count it, so firstly let me count here, how many candidates passed, 1, 2, 3, 4, 5 and here, 6. So there are 6 candidate, who passed. What about fail? 1, 2, 3 and 4. So there are four candidates, who failed. So now, this number, the number of candidate who passed and the number of candidate who failed? This is denoted by; say $n_1 + n_2$ and here n_2 . So, this category, this is here and 1 and this category this is going to be denoted by here, n_2 . And this number n_1 and n_2 , they are simply trying to represent, the number of candidates, in the category a 1 and in the category a 2. Or simply the number of candidates who fail, or the number of candidate, who belong to category a 1 or the number of candidates, that belongs to the category k, A2. So this n_1 and n_2 , they are simply trying to present, the number of units present in the category, a 1 and a 2. So the number of observation in a particular category, they are called as, 'Absolute Frequency'.

Refer slide time: (12:20)

Absolute and relative frequencies

The relative frequency of a_1 is $f_1 = \frac{n_1}{n_1 + n_2} = \frac{6}{10} = 0.6 = 60\%$

The relative frequency of a_2 is $f_2 = \frac{n_2}{n_1 + n_2} = \frac{4}{10} = 0.4 = 40\%$

This gives us information about the proportions of Pass and Fail persons in the test.

Now, one of the drawback, in absolute frequency is that, means, if I try to give that .Okay? There are 100 candidates who pass; there are 300 candidate who failed. But, you are not trying to see that, how many candidates appeared. So, in order to incorporate, this feature, we have a concept of relative frequency. for example, there are 6 candidate which passed ,there are four candidates which field, I can also say, that there are 6 candidate out of 10, who passed? And there are four candidates out of 10, who failed? So, I can denote, the relative frequency of the class a1, to be the number of persons in the class, divided by total numbers, in total number of observations, that are available and n_1 is the numbers in, a 1. And this is denoted as he here f_1 , f_1 is going to be N_1 upon, $N_1 + n_2$. So this is called, 'Relative Frequency'. And in this case, this number is 6 up on 10,because we have absolve 10 assets, so that is 0.6 or this can be called as,'60%'. So, I can say that there are 60%candidate, who passed? And similarly, the relative frequency of the second class a2, that is denoted by, f_2 . And the definition of f_2 , is the similar to the definition of f_1 , that is the total number of elements, in the category A2, total number of elements in category A1, divided by total number. So, that is going to be the number of candidate, who failed, that is

4/10. Which is here 0.4 or I can say 40%. So, this will give us an information, about the number of candidates, who pass or fail, with respect to the total number of candidate, who appeared in the examination. And this can also give us, the number in terms of percentage. So this is the basic idea, of the absolute and relative, frequencies. Now, next question comes, how to compute this absolute and relative frequencies in the R software, in order to compute ,this absolute frequency, first we need to define our data vector, the data vector as I said, in the earlier lecture, the data vector will consist all the numerical values and they are combined using the C operator. Right?

Refer slide time: (15:17)

Absolute and relative frequencies

data vector = $C(x_1, x_2, \dots, x_n)$

`table(data vector)` creates the absolute frequency of the data vector of the given data in the vector.

Enter data as **x**

`table(x)` # absolute frequencies

`table(x) / length(x)` # relative frequencies

Relative freq = $\frac{\text{Abs. freq}}{\text{Total freq}}$

→ `length(x)`

So, I can see here, net any data vector, that is going to be in the format of C and then here, the value X 1, X 2, up to here, X n. means, if I assume that there are n values, for example, in the earlier case there are 10 observation and is going to be 10 and soon. And so, after this the command to obtain the absolute frequency is table, table all in small letters. So, when I try to write down, this command table and inside the bracket, the data vector, then this will create the, absolute frequencies, of the data, which is given inside the arguments, under the vector, data vector. And suppose you want to find out the relative frequency, so now, if you try to understand, what is the relationship between frequency and relative frequency? Frequency means, absolute frequency. The relative frequency is, absolute frequency, divided by total frequency. Total frequency means, total number of observation. So, once I am trying to write down all the observation in a data vector, then whatever is the length of the data vector, that is your total frequency. So, this total frequency can be computed by the, by the command, length. a length and inside the brackets, you have to give the data vector, say here, X .so, now in case if you want to find out the absolute frequency, then I simply have to use the same command, table and inside the arguments, I have to give the data vector, say here X and I need to divide all the values by length of X. remember one thing, length of X is going to be a scalar value. But, now if you recall, when we did the division operation, of our data vector, with respect to a scalar, then we had learned, that when the division happens in each and

every element of the data vector. So, this value will give us, the absolute frequency divided by, total number of observation and this will give us the, information about, the relative frequency.

Refer slide time: (17:53)

Absolute and relative frequencies

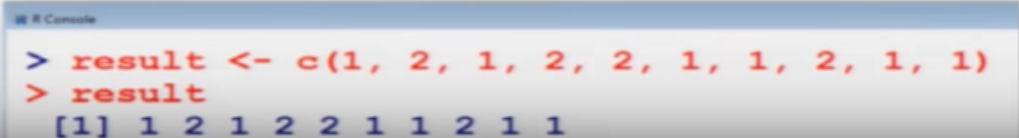
Results of 10 persons declared in two categories as Pass (P) and Fail (F) is categorised as 1 and 2 respectively.

$P \rightarrow 1$
 $F \rightarrow 2$

P	F	P	F	F	P	P	F	P	P
1	2	1	2	2	1	1	2	1	1

\rightarrow data.

```
> result <- c(1, 2, 1, 2, 2, 1, 1, 2, 1, 1)
> result
[1] 1 2 1 2 2 1 1 2 1 1
```



```
> result <- c(1, 2, 1, 2, 2, 1, 1, 2, 1, 1)
> result
[1] 1 2 1 2 2 1 1 2 1 1
```

So, now let me, try to take an example and try to show it, on the R software also. So, for example, I will take the same example, but now, I'm doing one thing more, means, earlier all the ten candidates, who were categorized in, two categories, as pass or fail? Now, I will try to assign them, an indicator value. Because, as we discussed in the earlier lecture, that unless and until, you try to assign a numerical value, to our data. We cannot operate any stat scale tool. so now I have here two values, one is a here pass and and there is, here fail. So, what I try to do here, for pass, I try to represent it by number one and for fail I try to represent it by number two. Once you do it, then this data P, that will be represented by one and this F will be represented by 2 and similarly, variable is your P that will be replaced by one and wherever is your F that will be replaced by two. So now, this is our data. And now, I need to type this data into a data vector, before I can, expose this to R software. so you can see here, that I have, created here a data vector like this and you can see here, this one is here this 1, this 2 here is this 2, this one is here this 1, this 2 here is this 2, this 2 is this, this one is this, this one is this, this 2 is this, this one is here one and this one here is one. And based on that, I have created my data vector here, result. So case you can see here, the outcome and here is the screenshot, of the same result, well I will show you on the arc console also, but before you let me, try to explain it. And then, it will be more convenient for me to show it, on the R software.

Refer slide time: (20:01)

Absolute and relative frequencies:

```
> table(result) # Absolute frequencies
```

```
result
```

```
1 2
```

```
6 4
```

1 → Category 1

2 → Category 2

6 → # of elements in
Category 1

4 → # of elements in Category 2

Absolute frequencies

```
R Console
> table(result)
result
1 2
6 4
> |
```

Now, I will simply use my table command and I write here, table and inside the argument I try the data vector name, that is a result, as soon as I write here, table and result inside the argument, it will give us this type of argument. So, first we try to understand, what is the meaning of this result? You can see here, there are four numbers here, one, two, six and here, 4. So, what this one is indicating? And what two is indicating, first of all. This one is indicating, the category. You will category one and then category two. And now, what this six and four are indicating, six is indicating, the number of elements in category one. And similarly, this four is indicating number of elements in category two. And this is nothing but, these are your absolute frequencies, six is the absolute frequency of category one that is, class1 and 4 is the absolute frequency of class 2. And similarly in case if you want to obtain the same result, with respect to the relative frequencies,

Refer slide time: (21:54)

Absolute and relative frequencies:

```
> table(result)/length(result) #Relative freq.
```

```
result
```

```
1 2
```

```
0.6 0.4
```

```
6 → n1
```

```
10 → n1 + n2
```

```
n1 = 6
```

```
n2 = 4
```

```
4 → n2
```

```
10 → n1 + n2
```

```
Relative frequencies of  
Categories 1 and 2 resp
```

```
Categories → 1, 2
```

```
R Console  
> table(result)/length(result)  
result  
1 2  
0.6 0.4
```

Then, what we have to do? I would write here, the same command table, inside the arguments the data vector, whose name is result? And I would divide it, by the length of the data vector, once you do it, you will get here an outcome like, this. What is this? Showing you, you can see here, there are 4 values, one, two, point six and point four. This one and two, as earlier, they are trying to show you, the categories. category one and category two and point six is actually, 6/10, six here is n₁ and 10 here is + 1 plus n₂, where you can see here N₁ is equal to 6 and n₂ is equal to here 4 and similarly this, 0.4 is 4 upon 10, which is here n₂ divided by N₁ plus n₂. So this, 0.6 and 0.4, these are the relative frequencies of categories, 1 & 2 respectively. And this is here, the screen shot that we are going to get, when we try to execute this thing on the R console. So, let me now, first come to R console and try to show you here, that how the things are happening, so first I will try to create here, a vector here, result and then I would try to show here, different things.

Refer slide time: (23:50)

```

> result <- c(1, 2, 1, 2, 2, 1, 1, 2, 1, 1)
> result
 [1] 1 2 1 2 2 1 1 2 1 1
>
> table(result)
result
1 2
6 4
>
> length(result)
 [1] 10
> table(result)/length(result)
result
 1  2
0.6 0.4
> |

```

So, you can see here, I have created here a, data vector, result like this. Now, I would try to create or try to find out here, the absolute frequencies, by using the command table, table and inside the argument, I will say, the data vector whose name is result, so you can get here, the similar output, that I discuss, this one, this is indicating the category one, this 2 this is indicating the category 2, this 4 this is indicating the number of elements in category 1 and this 4 is indicating the number of elements in category 2. But, definitely this is going to give you the result, in terms of absolute frequency, now in case if you want to have this result in terms of relative frequency, then first I will show you, that what is the, value of here, the length of result, this you can see here, this is 10 and you can count here 1, 2, 3, 4, 5, 6, 7, 8, 9 and here, 10. So there are 10 values in the data vector result, so now, if I try to write down table result, divided by the length of result, then the outcome comes out to be like this one. so you can see here, this one is representing the category 1 and this point 6, is trying to denote, the relative frequency of class 1 and this 2 is denoting the category 2 or class 2 and this 0.4 is greater noting the relative frequency of the class 2. So, this is how you can obtain the, the absolute and relative frequencies. And this, absolute and relative frequencies, you can see, there will be more prominent, when you have a qualitative variable. Now, give you an idea, what that, whatever we have done here, I would write to put them, in the right words. So, I had a set of data of 10 candidates, in terms of two categories, a1 and a 2 pass and fail. And we have found the number of persons, who are passed and number of persons, who are fail categories. So and then, I would try to or I would say that, I have rearranged the entire data sets into 2 groups. So, this arrangement of ungroup data, into group data.

Refer slide time: (26:38)

Frequency Distribution

- Arrangement of ungrouped data in the form of group is called frequency distribution of data.
- Classify the data into different classes by dividing the entire range of the values of variables into suitable number of groups called class.

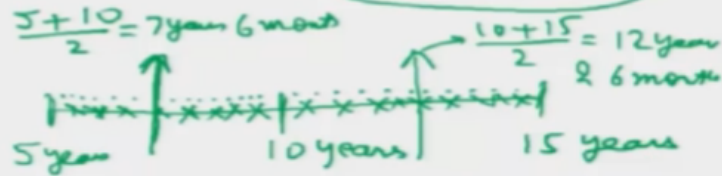
This arrangement of ungrouped data in the form of group, this is called the, 'Frequency Distribution of Data'. that's a standard terminology and we always call it, please create a frequency distribution, that means, you need to, group the data and then you have to condense the data, condense means, you can see that, the, the ten delayed Qatar values, have been condensed only into, two categories and they are based only on two value, six and four, when six is the frequency of class one and four is the frequency of class two. And now, what we try to do? Whatever is the data? this data is, condensed into different groups and for that, I try to, create different groups, for example, this group's a1 and a2, they are not coming from Sky, you are the one, who has created this group says, pass or fail. So, I will try to, divide the entire data into different groups and these, groups are called as. 'Classes'. So, the meaning of a class is simply a group and for the given set of data, we always try to create, suitable number of groups. well I'm saying here, suitable number of groups, well there is not a very hard and fast, rule to decide, how many groups can be, therefore that you have to, use your common sense and some basic information, about the experiment to decide, that how many groups, can really help you in getting the data or the information which is contained inside the data. Means, obviously if the number of group is, is too small or too large. Then possibly, it will be more difficult to handle the data. So, we need to have some suitable number of groups that we will try to see, with some lectures in the coming slides. Now, in every group, you have to define the boundaries, for example, in this case where we have only pass or fail, well there are no mathematical boundaries, but ,but they are categorized only by, by two, disk categories pass and fail. But, suppose you try to record the age or height, of say, some number of candidates ,then the age can be, 5 years, 7 years, 9years ,12 years, 20 years, 18 years, 21years, 30 years and so on. So, this ages can be defined into, can be defined in some groups, like as, five years to 10 years ,10 years to 15 years or this can be 0 to 10 years ,10 to 20years, 20 years to 30 years and so on. So, in this case, they will be representing our class, so whenever we are trying to define a class, there are going to be two values, one is the lower boundary of the class and upper boundary of the class

Refer slide time: (29:36)

Frequency Distribution

- Lower and upper boundary figures of a class are called the lower limit and upper limit respectively.

- Difference between the limits is called the width of the class or class interval.



- The value of variate lies in the middle of lower and upper limits.

And these are called as, 'Lower Limit and Upper Limit'. And when we are trying to, find the difference between, these two limits. The lower limit of the interval and the upper limit of the interval, then this is called as, 'Width', of the class or 'Class interval'. And when you are trying to define a class, there are two values, lower value and upper value. And this will be a sort of interval. Now, whatever is the value in the mid of this interval, this is called as, 'Mid Value'. And the advantage, of this mid value is that, when we are trying to group the data, the data will be scattered over the entire interval. But, we assume that, that entire set of data is going to be concentrated only at the mid value. So, in case, if you try to see here, what will really happen, suppose I try to take here, two intervals, say ages, five years, ages say here, ten years and ages, fifteen years. And suppose I try to collect the data, the age comes out to be seven years, it will come here. It comes out to be eight years that will come here. Now, it will come out to be nine years that will come here and so on. There will be many many observation in this interval. and similarly, all those edges which are lying between 10 years, 11 years, 12 and so on, up to 15 years, they will be lying in this interval, so all these values are lying over here, in this interval at, different locations. But, we assume that, once they are grouped, then all are going to lie in the mid of the interval. and the mid of the interval is, simply going to be five years, plus ten years, divided by two, is equal to seven years, six months. And similarly here, the mid value is, ten years, plus 15 years, divided by 2, this comes out to be 12 years and six months. So, this is what we assumed that, the value of the variant lies in the, middle of lower and upper limits.

Refer slide time: (32:14)

Frequency Distribution

- The number of observations in a particular class is called

absolute frequency or frequency.

$$\frac{\text{absolute freq}}{\text{Total \# of obs}}$$

- The number of observations in a particular class divided by total frequency is called relative frequency.

And whatever is the number of observation, which are lying inside a particular class, this is simply called a, 'Absolute Frequency or in simple words, we also call it as,' Frequency'. And when we are trying to divide, this absolute frequency, by the total number of observation in that class, then this value is called as, 'Relative Frequency', of that class.

Refer slide time: (32:51)

Frequency Distribution

- The cumulative frequency corresponding to any variate value is the number of observations less than or equal to that value.
- The cumulative frequency corresponding to a class is the total number of observations less than or equal to the upper limit of the class.

Now, there is another aspect, which is cumulative frequency. What is the cumulative frequency? As the word suggests, cumulative means, you are trying to accumulate; you are trying to add more and more. So the cumulative frequency is also defined, to a particular class and this is, defined for all the classes separately. So, the cumulative frequency, corresponding to any variate value, is the number of observation less than or equal to that value. And this cumulative frequency, corresponds to a class and what is the total number of observation, less than or equal to the upper limit of the class. What does this mean? Okay?

Refer slide time: (33:39)

Frequency Distribution

Example:

Following are the time taken (in seconds) by 20 participants in a race.

32, 35, 45, 83, 74, 55, 68, 38, 35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58.

min = 32
max = 84
width = 10 sec

The data is summarized in class intervals

30-41, 41-50, 51-60, 61-70, 71-80 and 81-90

A₁, A₂, A₃, A₄, A₅, A₆

Let me, try to take an example and try to illustrate all these things, then it will become more clear and easy to understand. Now, in this example, there are twenty participants, who participated in a race and time taken to complete, the race is recorded in seconds. So, this thirty two means that, the first participant took 32second. This thirty five means, the second participant took thirty five seconds. This forty five means, they are the third participant of the forty five seconds and so on. Now, what you have to look here, this is very important. And yeah, I would also request you that you please try to concentrate on this example, that how I'm trying to create class intervals and what are the steps which are involved. Because, in this lecture, I am going to explain, the example in detail and in the next lecture, I will try to implement the same example in R software. And one and when you are trying to implement it, it is important for you to, to do, the same steps in R software, which you are doing here .Okay? So, now looking at this data, first we try to see, what is the minimum value in this data? And what is the maximum value in this data? So, I can see here, that 32 is the minimum value and here, 84 is the maximum value. so the minimum value is 32 seconds and maximum value is 84 seconds. Now, looking at these two values minimum and maximum, I have to define here, the width of the class interval. And this width is, going to decide the number of intervals also. So by looking at this data, suppose suitably, I choose that the width of the interval should be of 10 seconds. so I try to, create here, different classes, like this, class one say here, a 1 this is consisting of 32 ,41 seconds. That means, this interval a 1, will contain all the values of the time, which are lying between 30 seconds and 41 seconds. And similarly, the finished class is a 2. Which will contain all the values between 41 seconds and 52nd? And similarly, I have here Class A three, a four, a five and here, a six. so I have created here, six classes and you can notice, that all this six classes, they can contain my all that data and that is another point, while creating these groups, that these groups and the limits have to be defined in such a way, such that the entire data can be accommodated, among these, groups. And now, in this group you will see here, this this 30 and 41.this 30 is the lower limit and 41 is

the upper limit. And similarly in this case, in the basics 81 is the lower limit and 91 is the upper limit. so now, I have, created the groups, in which the entire Decatur can be summarized.

Refer slide time: (37:27)

Frequency Distribution

Example (contd.):

Total no. of obs (n) = 20

Class intervals	Mid point	Absolute frequency (or frequency)	Relative Frequency	Cumulative Frequency
30 - 41	35.5	5	$5/20 = 0.25$	5
41 - 50	45.5	3	$3/20 = 0.15$	5+3 = 8
51 - 60	55.5	3	$3/20 = 0.15$	5+3+3 = 11
61 - 70	65.5	5	$5/20 = 0.25$	5+3+3+5 = 16
71 - 80	75.5	2	$2/20 = 0.10$	5+3+3+5+2 = 18
81 - 90	85.5	2	$2/20 = 0.10$	5+3+3+5+2+2 = 20
	Total	20	1	

Handwritten notes in the table:
 - Midpoint formula: $\frac{30+41}{2}$
 - Relative frequency denominator: Total obs
 - Cumulative frequency labels: A₁, A₂, A₃, A₄, A₅
 - Class interval labels: A₁, A₂, A₃, A₄, A₅
 - A note on the right: "abs. freq class"

Now, I will try to present it in some suitable tabular form, so that I can easily understand it. So, now you can see here, I have created a table. And the first column here, is class interval and in this interval you can see here, that I am trying to, give the same class interval, which I had denoted here as say, a 1, a 2, a 3, a 4, a 5 and a 6. And then in the next column, I am finding out the, midpoint. So, the midpoint of Class A one is thirty five point five, which is coming from 30 plus, 41 divided by two and so on. And similarly I have found the, midpoints of other classes, so we are going to now assume, that whatever data is, is a spread in the interval thirty to forty one, this is going to be concentrated, at thirty five point five and another point of difference will be thirty five point five, that I will say, that the data, that all the data in this interval will have the single value, at thirty five point five. And as I said, earlier and we had discussed in the earlier lecture, also once you try to group the data, the information, on the individual information is lost, but only the information, on the data is available, that this data belongs to which category or see which class. Now, after this, this is the third column, in which I am trying to count the absolute frequency or frequency, for example, here, this is the interval thirty to forty one. So, if you try to look at the data, how many values are lying between, thirty and forty one? So, this is the first value, this is here the second value and they say the third value, fourth value and you can see here, this is the fifty value here. And similarly if you try to find out here, how many values are lying between, 41 and 50, I can say here, this is the value for forty five one and then, if you move for the, forty two second and then, forty two once again here two three, so there are only three values, which are lying between 41 and 50, so you can see here, this is represented here, where I am trying to make us circle and similarly, I have found the, I have counted the number of observations in that particular category, I have done it manually and I have written here, so this fiber is indicating, that there are five values in the interval 61 to 70 and so on. Now, the total number of observation here, you can see, is 20. And this we try to denote by here and this is equal to here

20. So now, when I try to divide the absolute frequency by the total number of observation, I get here, the relative frequency in the next column. So, you can see here, this 5 is coming here, this 3 is coming here, this 3 is coming here and this is here the total number of observation. And once I try to divide it, I get here the values of relative frequencies. The advantage of relative frequency is that, all the relative frequencies, they will always belying between, 0 and 1 and so, they can be easily converted into percentages. Now, in the last column, I am trying to find out the, cumulative frequency and here, I would like to, once again explain you, that how these cumulative frequencies are found. Now, this is my here, I can rewrite here, class a 1 plus, a 2 plus, a 3 plus, a 4 plus, a 5 and class a 6. Now, I am trying to say, the total number of observation in the given set of data, in Class A 1 is 5 or this a 1 is, having the limit 32 ,41. So I can also say, that there are only 5 observation in the entire data set, where I am trying to make it circle there, there are only 5 observations, whose values are smaller than 41. Now, let me come to the, second category, a 2 this is going from 41 to 50. So you can see here, that that total number of observation, whose values are smaller than, 50. What are those things, 5 and the absolute frequency? Absolute frequency of group 1 and absolute frequency of group 2, of class 1 and class 2. So this is here, 5 and 3. And similarly, if you try to look at a 3, a 3 has a limit, 51 to 60. So, this is trying to give you, a value here, which is the sum of all the absolute frequencies up to the third class, which is 5 plus, 3 plus, 3, that is 11. and similarly for the a 4 this is trying to give you the total of all the absolute frequency, from class 1 to class 4 and similarly for the class 5, this is trying to give you, the sum of all the absolute frequencies up to the class, 5 and similarly this sixth and the last group is trying to give us, the sum of all the absolute frequencies up to the class 6 and this is obviously going to be the total number of observation. So, this is what we mean by, cumulative frequency. So, by looking at the value of the cumulative frequency, I can always find that how many values are smaller than this value. And now, the same thing I can just, make general. And now, I will say that, instead of here,

Refer slide time: (43:49)

Frequency Distribution

General, if there are k class intervals, n observations are divided into k class intervals a_1, a_2, \dots, a_k containing n_1, n_2, \dots, n_k observations respectively.

Relative frequency of j^{th} class : $f_j = \frac{n_j}{n}$ $j = 1, 2, \dots, k$

Frequency distribution:

Class interval (a_j)	a_1	a_2	...	a_k
Absolute frequency (n_j)	n_1	n_2	...	n_k
Relative frequency (f_j)	f_1	f_2	...	f_k

6 class interval, I have here, K class interval, in general. And there are total number of observation, are here N. And this observations are divided into K class interval. And this class intervals are denoted by a 1, a 2, a K, in such a way, such that class a 1 contains, n1 observation, a 2 contains N2 observation and a K contains here, NK observations respectively. So, obviously if you want to find out the relative frequency of the J th class, this will be here, the number of observation in J th class divided by the total number of observations. and j will goes from 1, 2 up to here k and now all this information, can be combined together, in this format class interval here, I can write down a 1, a 2, a K, then the absolute frequency N 1, n 2, NK, relative frequency F 1, F 2, F K and if required, I can also add the information on the cumulative frequency. So, this table, what we have drawn here, this is called as, 'Frequency Table', or The Frequency Distribution.' why we call it distribution? Because, we are trying to see, how the values of a variable are distributed. So, now in this lecture, if you try to see, I have taken an example and then, based on that example, I have tried to give you, the different definition, concepts and how the things are implemented. But, whatever I have done, that is manually. Now, in the next lecture, I will continue with the same example, but then, I will try to show you, that how the things can be implemented over the R software. So, you practice it, you try to learn it, you try to understand it and we will meet in the next lecture, till then, Good bye.