

Essentials of Data Science with R Software - 1
Professor Shalabh
Department of Mathematics & Statistics
Indian Institute of Technology Kanpur
Lecture No. 29
Quantiles and Tschebyshev's Inequality

Hello friends, welcome to the course Essentials of Data Science with R software 1 in which we are trying to understand the basic concepts of probability theory and statistical inference. So, now, you can see up to now, we have understood different types of characteristics of the data and we have learned that how they can be defined on the basis of theory and how they can be implemented in practice based on a sample of data and using the software.

So, we will continue on the same lines in this lecture also and we will try to learn the concept of quantile and after that, I will conclude this lecture by 1 more Tschebyshev inequality. So, first of all let me try to give you an idea of what is the meaning or what is the utility of quantiles and Tschebyshev inequality.

Have you heard many times that in certain competitive examination, there is a condition that any student who has got the marks more than the say eightieth percentile of the board then the person is eligible to appear in the examination. What is the meaning of this? One thing I would say do not get confused with the 80 percentage; 80 percent, 80 percentage they are different than eightieth percentile. Percentage that we know how to calculate, means out of 100 what will happen. But this percentile is a different concept.

So, the first question come why do you use such a percentile approach and how are you going to compute it and how you can make it more general. Now, as a hypothetical situation we can assume that suppose there are two boards of examinations say class tenth or class twelfth means any1 you can choose suppose we take the class twelfth board. So, we know at least in our country, various states have boards what we call a state board, UP board, MP board, West Bengal board and then we have some Central Board also like CBSE etc..

So, these are the examination bodies which conduct the examination of the same class level in different places and suppose I takes, suppose there are two boards board 1 and board 2. Sometimes we hear or we people say that this board gives the mark not so easily and in this board it is very easy to get these marks What is the meaning of this or sometimes people say that okay the marking in the say, for example board number 1 is harder than the marking in the board number 2, I do not know whether this is true or not, but just for the sake of

understanding. For a while you assume that suppose, this statement is true that board 1 has the marking scheme which is harder and suppose both to has a marking scheme which is not so hard, that is just lenient.

Now, what will happen suppose there is a question paper in suppose say mathematics. Now out of 100 marks, a student in the board 1 gets only say 40 percent of marks and the student in the board 2 get 70 percent marks, means out of 100 they get 40 and 70 marks, respectively in board 1 and board 2. Now can you really compare them. You are saying that the board number 1 has a hard marking scheme that it does not give the mark so easily and it will take out the mark even for a single mistake but in the board 2, the marking scheme is lenient, they will not bother about for example, say these smaller mistakes.

So, now can you compare the two students having the 40 percent marks and 70 percent marks? Certainly not because it is possible that the student in the board 1 is far much better than the student in the board in the board 2, as far as the concepts are concerned. So, the question is this you cannot compare by percentage. So, what to do one simple option is this when the students are appearing board 1 and in board 2, then the marking scheme within board.

One is homogeneous and within board 2 that is also homogeneous, they are similar, they are not comparable when they are trying to compare according to the boards. So, now, inside the board 1 if there are 2 students somebody has got 40 percent and somebody has got 60 percent I can say very easily that the student with the 60 percent mark is better than the student with 40 percent marks and same thing is true for the board 2 also, but when we are trying to compare the two boards, then they are not comparable.

So, what we try to do, we try to take the marks of the board 1 and we try to divide the minimum and maximum marks as 0 or and 100 and then we try to create the probability distribution and we try to create 100 parts of the distribution of marks and the same thing I try to do in the board 2 also that minimum and maximum marks they are equated as say 0 and 100 and then we try to make 100 partitions of the distribution of the marks in the board 2.

Now, what will happen that in the suppose in the eightieth partition of marks in board 1 suppose that comes out to be only say 50 and eightieth partition in the board 2 comes out to be as suppose here 90. So, that means, within that group whatever are the students at the

eightieth part, eightieth positions in their respective board, they are comparable that is a way how we can think of comparison.

So, this concept is called actually quantile and in particular this case when we are trying to create 100 partitions these partitions are called as percentile. So, this is how this quantile concepts concept helps us in data sciences when we, whenever we are trying to compare the values in that two datasets. People also try to compute different types of probabilities using the quantiles of a distribution.

For example, if somebody wants to know what is the probability that the total number of people who are visiting our shopping website are in the top 5 percent, what is the probability? So, then that means, they have to simply compute the ninety fifth percentile and then they have to see that how many people are lying in the interval from ninety fifth onwards to 100 percentile. So, this is a very useful concept. So, that is what we are going to learn here number 1.

Number 2, the second concept which I am going to deal in this lecture is about the Tschebyshev inequality. Whenever you are trying to find out the probability of an event, you always try to approximate it using some probability function it can be for a discrete random variable or a continuous random variable. But definitely you are trying to compute the probability on the basis of some given set of observations and this probability may not be 100 percent accurate.

In that case you would like to see, ok, what is the minimum value of the probability beyond which we cannot go or means, if you try to find out the probability of certain event and if you try to find out that the probability cannot be lower than this, then this types of, these types of statements they really help us in data sciences when we are trying to work with the real data. Because ultimately your objective is that you want to get the correct statistical inference. So, Tschebyshev inequality helps in finding out such probabilities. So, if you can find out the probabilities using the Tschebyshev inequality, suppose this comes out to be 0.3 and your estimated probability is suppose 0.5. So, you can see that, I am trying to compute it as 0.5 and if I try to take another sample possibility that may come to 0.4 also, but in any case that cannot go beyond 0.3.

So, these types of statements are drawn using the Tschebyschev inequality, which is a very simple thing, but it is I personally believe that it is important for you to understand. So, let us try to begin our lecture and try to first understand the quantile concept.

(Refer Slide Time: 10:34)

Quantiles: CDF

We define quantiles in terms of the distribution function.

The value x_p for which the cumulative distribution function is

$$F(x_p) = p \quad (0 \leq p \leq 1)$$

is called the p-quantile.

$\Rightarrow x_p$ is the value which divides the CDF into two parts:

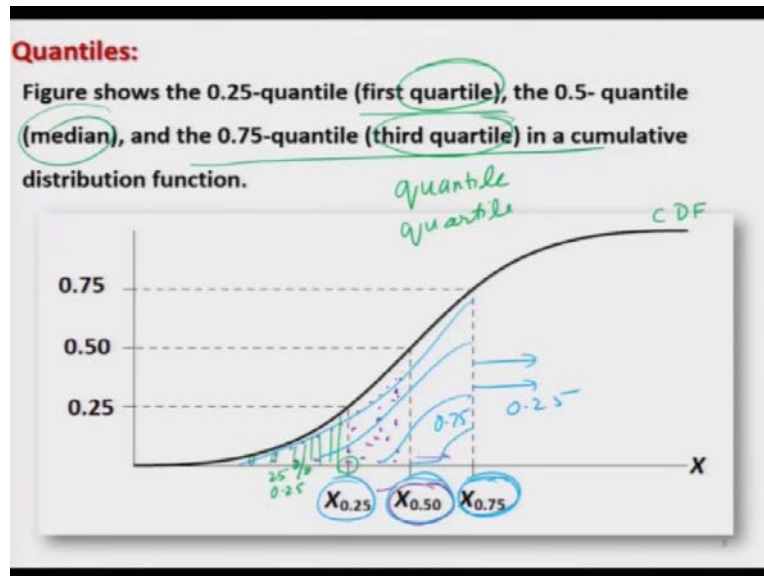
- the probability of observing a value left of x_p is p , whereas
- the probability of observing a value right of x_p is $1 - p$.

For example, the 0.25-quantile $x_{0.25}$ describes the x-value for which the probability of observing $x_{0.25}$ or any smaller value is 0.25.

So, we define the quantiles in terms of distribution function and this was a cumulative distribution function which you had indicated as CDF. So, suppose there is a value x_p , so this value x_p for which the CDF is given by $F(x_p)$ is equal to p , where p is some value lying between 0 and 1 is called a p th quantile or a p quantile. So, what is this value x_p doing here, x_p is the value which divide the CDF into two parts.

The first part is the probability of observing a value left of x_p is p , whereas the second part is the probability of observing a value right of x_p is $1 - p$. So for example, if you say we have the 0.25 quantile which will be indicated by x subscript 0.25. This will describe the x value for which the probability of observing this value $x_{0.25}$ or any smaller value is 0.25. So, that is as simple as that.

(Refer Slide Time: 12:02)



So, for example, if you try to see here suppose this is the CDF of some random variable here X and if you try to see here this $X_{0.25}$ is the value which is trying to divide it such that this area which I can highlight here this is actually 25 percent of the total area or it is 0.25 and similarly, if you try to divide this total part into two equal parts, so the quickest will be $X_{0.50}$, so this will be the area which is trying to cover this area.

So, this will be your here the value of X which is trying to divide the CDF into two parts as that the area below this value and above this value they are the same say 0.5 0.5 and similarly, if you try to take here say $X_{0.75}$ then that means $X_{0.75}$ is the value which is trying to divide the CDF into two parts such that the area below this value which is up to here, this is 0.75 and area beyond this thing this is 0.25.

So, this I can call it $X_{0.25}$ is my 0.25 quantile and $X_{0.5}$ is my 0.5 quantile and 0.75 is my 0.75 quantile or they are called as say here, first quartile, second quartile or the third quartiles of the CDF and actually this second quartile is called the median that you know that you have learned in your undergraduate classes or even the high school or class twelfth also that what is called the median.

Median is the value which divides the total frequency into two equal parts and you can see here this median is the value which is trying to divide the total area of a CDF into two equal parts. So, the second quartile is nothing but your median. So, first quartile, second quartile and third quartile, you have to be careful I am using here two sounds and in two words, one here is quantile and another here is quartile.

So, this is what I have written here, this is first quartile and this is your second quartile which is median and third quartile, which is here 0.75 quantile. Just be careful with my sign sound and voice.

(Refer Slide Time: 14:57)

Quantiles:

25% Quantile: Splits the data into two parts such that at least 25% of the values are less than or equal to quantile and at least 75% of the values are greater than or equal to the quantile.

50% Quantile: Splits the data into two parts such that at least 50% of the values are less than or equal to quantile and at least 50% of the values are greater than or equal to the quantile.

50% Quantile: Median

4

So, this is what I have tried to define here. The same quantiles they can also be called as 25 percent quantile, 50 percent quantile and so 25 percent quantile that splits the data into two parts, such that at least 25 percent of the values are less than or equal to the quantiles value and at least 75 percent of the values are greater than or equal to the quantile value and similarly, the 50 percent quantile is split the data into two part such that at least 50 percent of the values of less than or equal to the quantile and at least 50 percent of the values are greater than or equal to the quantile and this 50 percent quantile is actually called as median.

(Refer Slide Time: 15:42)

Quantiles:

$(\alpha \times 100)\%$ quantile: Value which divides the data in proportions of $(\alpha \times 100)\%$ and $(1 - \alpha) \times 100\%$ such that at least $(\alpha \times 100)\%$ of the values are less than or equal to the quantile and at least $(1 - \alpha) \times 100\%$ of the values are greater than or equal to the quantile.

So, now in general, I can define here that α into 100 percent quantiles is the value which divides the data in proportions of α into 100 percent and $1 - \alpha$ into 100 percent such that at least α into 100 percent of the values are less than or equal to the quantiles and at least $1 - \alpha$ into 100 percent of the values are greater than or equal to the quantile and if you are getting confused with this language, I will say simply try to put here α is equal to 0.25, 0.5, 0.75 and try to understand the interpretation what I have given here in this car in this curve, in this graph.

(Refer Slide Time: 16:28)

Quantiles:

Quartiles

The values which divide the given data into four equal parts, say, Q_1, Q_2, Q_3, Q_4

Q_1 : First quartile which has 25% of the observations.

Q_2 : Second quartile which has 50% of the observations – median.

Q_3 : Third quartile which has 75% of the observations.

Q_4 : Fourth quartile which has 100% of the observations.

So, whenever we are trying to divide the given data or the total frequency into four equal parts, they are called as quartile and they are indicated by Q_1 , Q_2 , Q_3 , Q_4 that standard symbol in the Applied Statistics. So, Q_1 is the first quartile, which has 25 percent of the observation Q_2 is the second quartile, which is 50 percent of the observation which is called as median and Q_3 is the third quartile, which has 75 percent of the observations and Q_4 is the fourth quartile, which has 100 percent of the observation and you have seen that these types of terminology are very well used in real data application people always want to know that how the data is partitioned into different values.

For example, you can see that whenever we have some sort of special shopping days, which are advertised heavily by some shopping websites, what you can see here that if there are suppose, say 1 million customers which are going to access the website suppose in that 3 days of time, then you see that as soon as the sale begins the number of customers which are trying to access the website in the beginning they are much higher than those who are in the middle part.

So, now, in this case, do you think that the number of customers in the first couple of hours they are going to be the same as in the last couple of hours? No. That is not going to be a symmetric distribution and under this situation, the concept of quantiles helps us in making different types of comparison.

(Refer Slide Time: 18:16)

Quantiles:
Deciles

The values which divide the given data into ten equal parts, say,

D_1, D_2, \dots, D_{10}

D_1 : First decile which has 10% of the observations.
 D_2 : Second decile which has 20% of the observations.
 D_5 : Fifth decile which has 50% of the observations – median.
 D_9 : Ninth decile which has 90% of the observations.

So, similarly, if you try to divide the total frequency into a 10 equal parts then this is called as deciles and they are indicated by D_1 , D_2, \dots, D_{10} . So, D_1 is the first day decile that has 10

percent of the observation, similarly D_2 has 20 percent of the observation similarly, D_5 has 50 percent of the observation and this is called as fifth decile and ninth decile which has 90 percent of the observations.

(Refer Slide Time: 18:48)

Quantiles:
Percentiles

The values which divide the given data into hundred equal parts, say, P_1, P_2, \dots, P_{100}

P_1 : First percentile which has 1% of the observations.
 P_2 : Second percentile which has 2% of the observations.
 P_{50} : Fiftieth percentile which has 50% of the observations - median.
 P_{90} : Ninetieth percentile which has 90% of the observations.

And similarly, if you want to divide the total frequency of the given data into 100 equal parts P_1, P_2, \dots , say P_{100} then this is called as percentile that is the same thing about a which I took the example in the beginning. So, P_1 is going to be the first percentile which has just 1 percent of the total observations, P_2 is the second percentile which has just 2 percent of the total observation.

Similarly, P_{50} , you can understand this the fiftieth percentile which has 50 percent of the observation and that is called as median. So, similarly, the ninetieth percentile is indicated P_{90} and it has 90 percent of the observation. So, these concepts helps you a lot but the main thing here is this how to compute them in the R software.

(Refer Slide Time: 19:38)

Quantiles:
R Command :

quantile(x, ...)

quantile(x, probs =, type =, ...)

Arguments

x numeric vector whose sample quantiles are wanted,
probs numeric vector of probabilities with values in [0, 1].
type an integer between 1 and 9 selecting one of the nine
 quantile algorithms

So, computation of these quantities in the R software is very easy. For that we have a command here `quantile`, `q u a n t i l e` and inside the parenthesis you have you have to give the data vector `x` and there are several options that you can give. But I will try to take here only one option here `p r o b s` `probs` to illustrate that how you can compute different types of percentile and in this case, for example, `x` is going to be the numeric vector or the data vector for which we want to compute the quantiles and `probs` is `p r o b s` that is going to indicate the probabilities means you want to partition the total area under different types of partitions.

And these partitions may be of equal size or they may be of unequal size like as you can see here, if the distribution is like this, I want this point and this point. So, the since the area under the curve, that is actually the probability density function or probability curve, so this area is always going to be 1 and this 1 area is going to be divided into say the required number of partition according to our choice. So, this `probs` function is going to help in fixing those sizes.

So, this process is a numeric vector of probabilities whose values will always be between 0 and 1 and `type` is an integer between 1 and 9 selecting 1 of the 9 quantile algorithms. Why? Because you see whenever you are trying to compute such partition, they are not computed by a simple methodology, but some algorithm is used to compute these partitions and different people have given different types of these algorithms.

(Refer Slide Time: 21:40)

Quantiles:
R Command :
 R offers nine different ways to obtain quantiles, each of which is chosen by the `type` argument.

Type 1 : Inverse of empirical distribution function.
Type 2 : Similar to type 1 but with averaging at discontinuities.
Type 3 : Nearest even order statistic.

10

So, these all these algorithms are going to certainly give you different values, but definitely these values are not going to differ much. So, in practice even unless and until you have a specific requirement, you can use any of the algorithm, for example in case if you give here type 1, the type 1 is based on the inverse of empirical distribution function type 2 is based on say type 1, but it is a little bit different that it is related to the averaging at discontinuities. Type 3 is the based on nearest even orders statistic and so on.

(Refer Slide Time: 22:19)

Quantiles:
Example
 Height of 50 persons in centimetres are recorded as follows:

```
166,125,130,142,147,159,159,147,165,156,149,164,137,166,135,142,
133,136,127,143,165,121,142,148,158,146,154,157,124,125,158,159,
164,143,154,152,141,164,131,152,152,161,143,143,139,131,125,145,
140,163
```

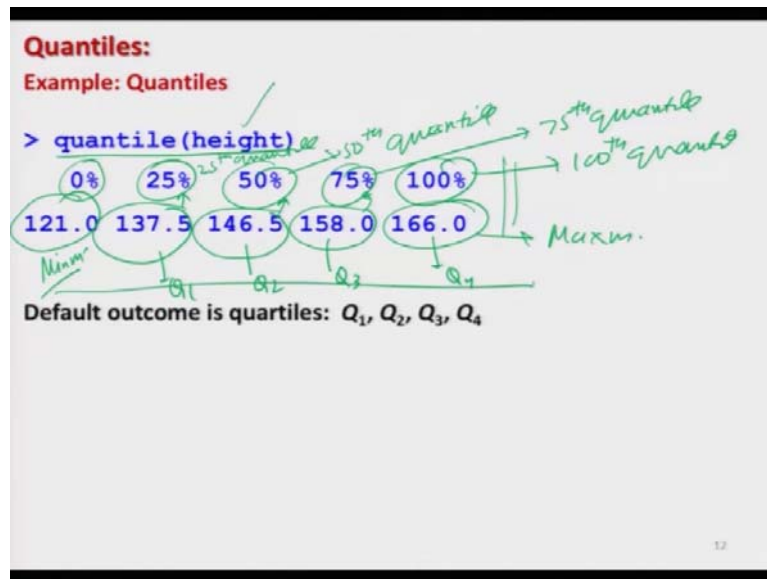
```
> height = c(166,125,130,142,147,159,159,147,
165,156,149,164,137,166,135,142,133,136,127,143,
165,121,142,148,158,146,154,157,124,125,158,159,
164,143,154,152,141,164,131,152,152,161,143,143,
139,131,125,145,140,163)
```

11

So, I am not going into that detail, but my objective here is to show you how to compute this partitions. So, suppose let me try to take an example here to illustrate that how these things are working because they are more useful in application. So, their application is more important for you to learn.

Suppose I take here only 50 observations and these are the observations on the heights offer 50 persons and they are recorded in centimetre, well I am taking here only 50 observation because if I try to take more observation, it is not possible for me to show you clearly on the screen, on this slide. But yes, this is small example will give you more ideas. So, I tried to store these values in a data vector your height.

(Refer Slide Time: 23:08)

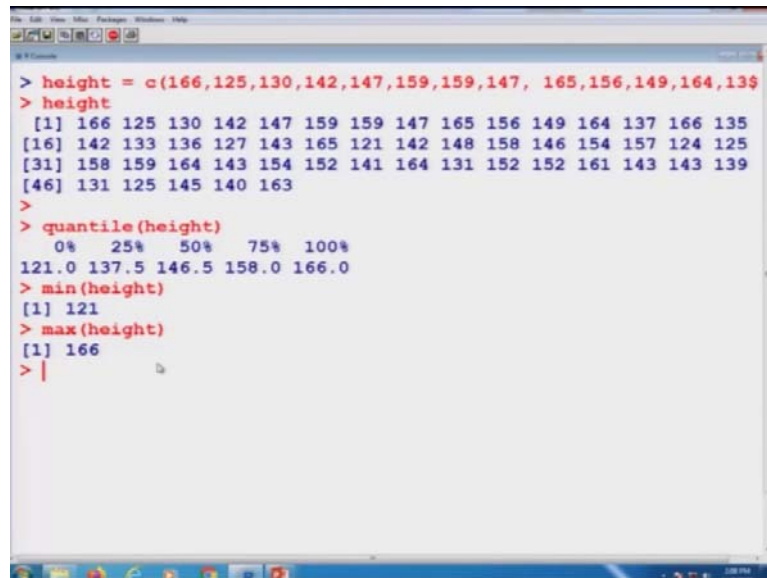


And then I tried to find out here the quantile height on the R console. Now, if you try to see how are you going to interpret it. As soon as you give this function over here, you will get an outcome like this one. So, this will start from 0th quantile and then it will come to twenty fifth quantile, quantile and this will be the fiftieth quantile and this is the seventy fifth quantile and this is here the 100 quantile.

So, this is going to be the 100 quantile, which is going to be the maximum value, because I will try to show you on the R console and this 0 percent that is going to be give you the minimum value also that it is the same as minimum value and this 130 7.5, this is the twenty fifth percentile, 146.5 is the fiftieth percentile, this 150.0 is the seventy fifth percentile.

So they are trying to give you a give us the value of Q₁, Q₂, Q₃, Q₄, so this is the value of Q₁, this the value of Q₂, this is the value of your Q₃ and this is the value of here I will try to show you on the R console also, but, first we try to have a look and then I will try to show you means other things. So, let me try to copy here this data that will save my time.

(Refer Slide Time: 24:48)



```
> height = c(166,125,130,142,147,159,159,147, 165,156,149,164,135)
> height
 [1] 166 125 130 142 147 159 159 147 165 156 149 164 137 166 135
[16] 142 133 136 127 143 165 121 142 148 158 146 154 157 124 125
[31] 158 159 164 143 154 152 141 164 131 152 152 161 143 143 139
[46] 131 125 145 140 163
>
> quantile(height)
 0%   25%   50%   75%  100%
121.0 137.5 146.5 158.0 166.0
> min(height)
[1] 121
> max(height)
[1] 166
> |
```


So, this is my here data and you can see here this data is like this, so now we try to operate the quantile function on this data vector height. So you can see here, these are here the height. So, this is going to be the twenty fifth quantile, which is the first quartile having the value and 37.5 this is the fiftieth quantile or this is the second quartile whose value is 140.5 this is here, the seventy fifth quantile or the third quartile whose value is 158 and this is here the hundredth percentile or this is the fourth quartile, whose value is 166.

So, as I told you that, this is 0 and 100 percent quantiles are the minimum and maximum value, so let us try to verify this thing. So, if you try to see here, the minimum value of this data vector height is here 121, which is same as here like this one and the maximum value of this data vector is 166, which is here, this value. So, you can see here that this quantile function has divided the range of 121, 166 into 100 equal parts. So, but now definitely you would have one more option that you do not need these partitions to be of the equal size. So, now how to get it done.

(Refer Slide Time; 26:24)

Quantiles:
Example: Quartiles Q_1, Q_2, Q_3, Q_4

```
> probs = seq(0, 1, 0.25) #probs for quartiles
> probs
[1] 0.00 0.25 0.50 0.75 1.00
```

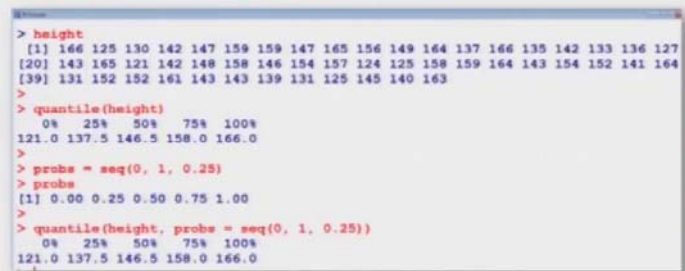


```
> quantile(height, probs = seq(0, 1, 0.25))
 0%  25%  50%  75% 100%
121.0 137.5 146.5 158.0 166.0
```

Same as earlier using `quantile` function.

13

Quantiles:
Example: Quartiles Q_1, Q_2, Q_3, Q_4



```
> height
[1] 166 125 130 142 147 159 159 147 165 156 149 164 137 166 135 142 133 136 127
[20] 143 165 121 142 148 158 146 154 157 124 125 158 159 164 143 154 152 141 164
[39] 131 152 152 161 143 143 139 131 125 145 140 163
>
> quantile(height)
 0%  25%  50%  75% 100%
121.0 137.5 146.5 158.0 166.0
>
> probs = seq(0, 1, 0.25)
> probs
[1] 0.00 0.25 0.50 0.75 1.00
>
> quantile(height, probs = seq(0, 1, 0.25))
 0%  25%  50%  75% 100%
121.0 137.5 146.5 158.0 166.0
```

14

So, for that we are going to use the option here, say `probs`. So, `probs` say for example, if I try to take care of sequence 0 to 1 at an interval of 0.25. So, a sequence 0 to 1 means this starting from 0 to 1 and this is here by 0.25. So, definitely this sequence is going to generate a values, a set of values like this 0, 0.25, 0.50 0.75 and 1. So, now, if you try to put here the quantiles of the same data set heights, but if you try to say `probs` equal to sequence 0 to 1, by 0.25

So, you can see here what are you trying to do, you are trying to divide the entire data values into 25 percent 50 percent 75 percent that means, you are simply trying to find out the quartile. So, you can see here that in case if you try to operate it on the R console, you will get here the same value and let me and if you try to compare it with here these values where

you have given only the variable name height without the probs function, you are getting here the same value.

So, this is what I wanted to show you, so that you are confident that when we are trying to use the probs function probs function is simply trying to divide the entire data into different partitions, when you want to have equal size partition you can give it or you want to have only four partitions you can give the values to be here 0.25, 0.5, 0.75 and 1 and in case if you want to change it, you can always change it. So for example, this is a screenshot of the same operation, I will try to show you on the R console also.

(Refer Slide Time: 28:11)

```

Quantiles:
Example: Deciles  $D_1, D_2, \dots, D_{10}$ 
> probs = seq(0, 1, 0.1) # probs for deciles
> probs
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

> quantile(height, probs = seq(0, 1, 0.10))
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
121.0 126.8 134.6 140.7 143.0 146.5 152.0 156.3 159.0 164.0 166.0
1  D1 D2 D3 D4

```

Need to change the probs function only.

```

Quantiles:
Example: Deciles  $D_1, D_2, \dots, D_{10}$ 

> height
[1] 166 125 130 142 147 159 159 147 165 156 149 164 137 166 135 142 133 136 127
[20] 143 165 121 142 148 158 146 154 157 124 125 158 159 164 143 154 152 141 164
[39] 131 152 152 161 143 143 139 131 125 145 140 163

>
> probs = seq(0, 1, 0.1)
> probs
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

>
> quantile(height, probs = seq(0, 1, 0.10))
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
121.0 126.8 134.6 140.7 143.0 146.5 152.0 156.3 159.0 164.0 166.0
>

```

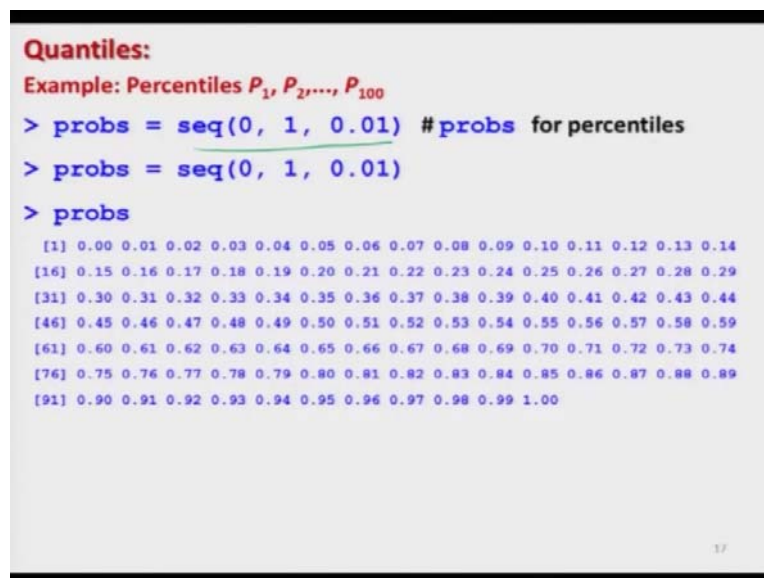
But before that means if I try to change my probs function and if I try to take it here sequence from 0 to 1 by 0.1 that means you are going to get here 10 values starting from 0.0 to 1 and at

an interval of 0.1. So, the values will be 0.0, 0.1, 0.2 upto 0.9 and then finally 1. So, now if you try to use this probs function in the quantiles function, then you are giving here probs equal to sequence 0 to 1 by 0.1 and now you can see here the outcome, outcome will look like this.

So, what are these things, they are trying to divide the total values into 10 partitions and this is called as deciles and they are the values of D_1, D_2, \dots, D_{10} . For example, this is here the value of your this is your 0 percentile, this is the value of 10 percent say here quantile, I can say that at 0 whatever you want, you can call it because it is a 0 quartile or 0th percentile or 0th quantile but four to 10 percent this is the value of D_1 , which is here 126.8.

This is the value of your, say here D_2 , which is 134.6. This is the value of here D_3 which is the value your third decile and this is the value of your for third decile 143 and similarly here you can see here or you have obtained all that deciles or you have partitioned the total frequency into same size but having 10 partitions. So you can see here, this is the screenshot.

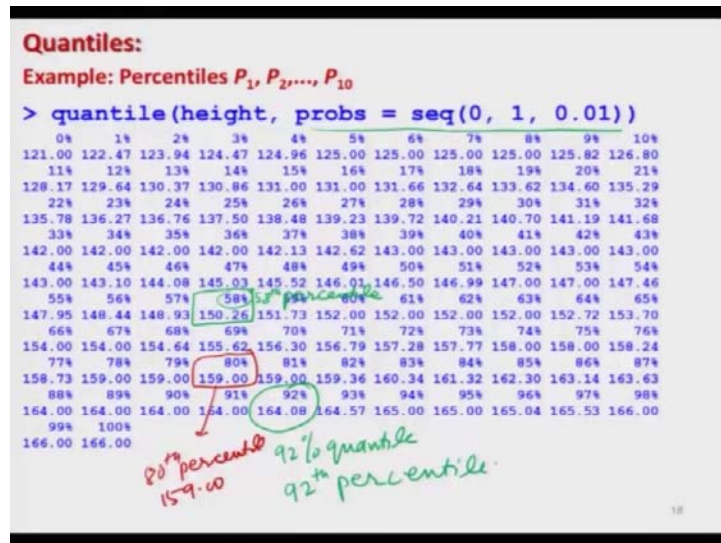
(Refer Slide Time: 30:02)



```
Quantiles:
Example: Percentiles  $P_1, P_2, \dots, P_{100}$ 
> probs = seq(0, 1, 0.01) #probs for percentiles
> probs = seq(0, 1, 0.01)
> probs
 [1] 0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10 0.11 0.12 0.13 0.14
[16] 0.15 0.16 0.17 0.18 0.19 0.20 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29
[31] 0.30 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 0.40 0.41 0.42 0.43 0.44
[46] 0.45 0.46 0.47 0.48 0.49 0.50 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59
[61] 0.60 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69 0.70 0.71 0.72 0.73 0.74
[76] 0.75 0.76 0.77 0.78 0.79 0.80 0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89
[91] 0.90 0.91 0.92 0.93 0.94 0.95 0.96 0.97 0.98 0.99 1.00
```

And similarly if you try to use here the option probs equal to sequence 0 to 1 at an interval of 0.01 the case will create 100 values and in case if you try to use this probs function inside the quantile function, what do we expect that will give you the values of percentile. So, you can see here.

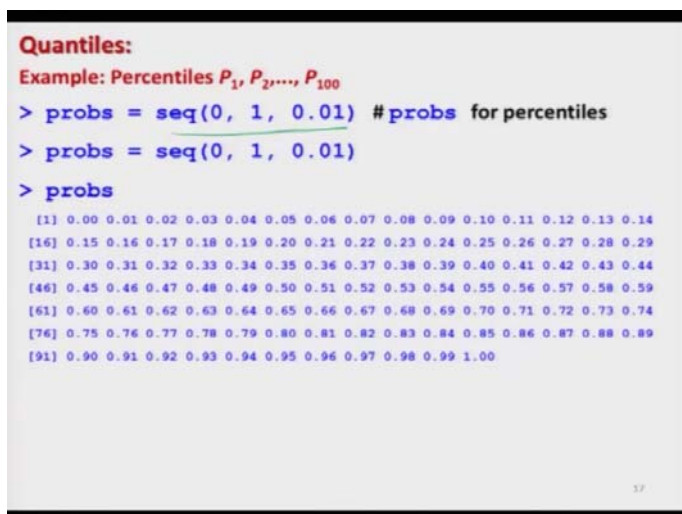
(Refer Slide Time: 30:23)



Now, in case if I tried to use here the quantile function with the probs equal to sequence from 0 to 1 by 0.01, then it is going to give you here this type of outcome here you can see, for example, if I try to see here, suppose, this value. What is this thing? This is the 58th quantile or this is 58th percentile and similarly, if you try to look at here, this is 92th percent quantile and this is the 92th percentile.

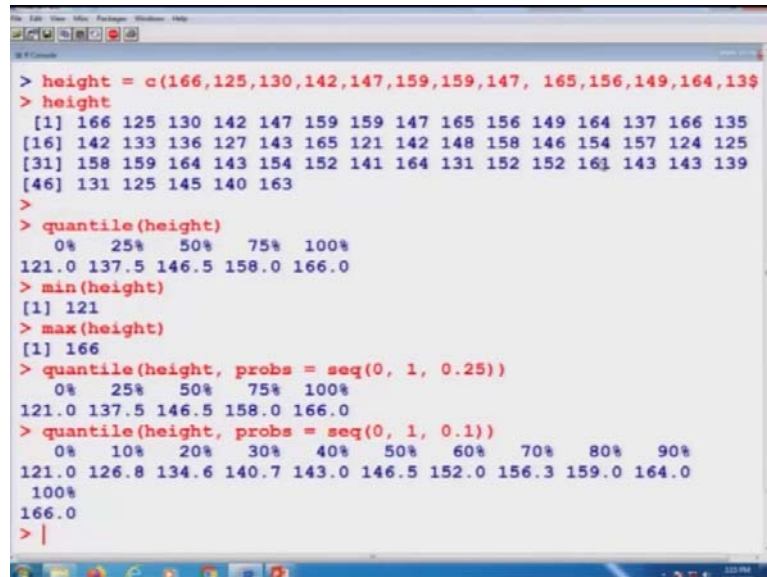
So, now you can see here, once you say that during an exam, anybody who is going to get suppose, more than the marks of eightieth percentile that is 20 is going to be eligible to appear in the exam that means, I simply have to find out here, this value so 80th percentile is just 159.00 marks. So, it means any student who has got marks more than 159 that person is eligible to appear in the exam.

(Refer Slide Time: 31:44)



And you can see here, this is the screenshot of the same operation, what I have just shown you and now, let me try to show you these outcomes what we have obtained here, for deciles and percentiles, etc., whether they really hold or not. So, now, I tried to find out here, the same function here, see here quantile.

(Refer Slide Time: 32:11)



```
> height = c(166,125,130,142,147,159,159,147, 165,156,149,164,135)
> height
[1] 166 125 130 142 147 159 159 147 165 156 149 164 137 166 135
[16] 142 133 136 127 143 165 121 142 148 158 146 154 157 124 125
[31] 158 159 164 143 154 152 141 164 131 152 152 161 143 143 139
[46] 131 125 145 140 163
>
> quantile(height)
 0%  25%  50%  75% 100%
121.0 137.5 146.5 158.0 166.0
> min(height)
[1] 121
> max(height)
[1] 166
> quantile(height, probs = seq(0, 1, 0.25))
 0%  25%  50%  75% 100%
121.0 137.5 146.5 158.0 166.0
> quantile(height, probs = seq(0, 1, 0.1))
 0%  10%  20%  30%  40%  50%  60%  70%  80%  90%
121.0 126.8 134.6 140.7 143.0 146.5 152.0 156.3 159.0 164.0
100%
166.0
> |
```

But by using the probs command, you can see here that this was you're here quantile height that is giving you four quartiles and now you are taking here the probs function. So, this is again giving you the quantiles. Now, similarly, if you want to find out here, say here, you want to divide it into 10 equal parts. So, you have to simply make here, the quantiles of high probs equal to starting from 0 to 1 at an interval of 0.1 and this will give you here 10 values, which are D1, D2 D10.

(Refer Slide Time: 32:48)

```

> quantile(height, probs = seq(0, 1, 0.33))
0% 33% 66% 99%
121 142 154 166
> |

```

0%	1%	2%	3%	4%	5%	6%	7%	8%
121.00	122.47	123.94	124.47	124.96	125.00	125.00	125.00	125.00
9%	10%	11%	12%	13%	14%	15%	16%	17%
125.82	126.80	128.17	129.64	130.37	130.86	131.00	131.00	131.66
18%	19%	20%	21%	22%	23%	24%	25%	26%
132.64	133.62	134.60	135.29	135.78	136.27	136.76	137.50	138.48
27%	28%	29%	30%	31%	32%	33%	34%	35%
139.23	139.72	140.21	140.70	141.19	141.68	142.00	142.00	142.00
36%	37%	38%	39%	40%	41%	42%	43%	44%
142.00	142.13	142.62	143.00	143.00	143.00	143.00	143.00	143.00
45%	46%	47%	48%	49%	50%	51%	52%	53%
143.10	144.08	145.03	145.52	146.01	146.50	146.99	147.00	147.00
54%	55%	56%	57%	58%	59%	60%	61%	62%
147.46	147.95	148.44	148.93	150.26	151.73	152.00	152.00	152.00
63%	64%	65%	66%	67%	68%	69%	70%	71%
152.00	152.72	153.70	154.00	154.00	154.64	155.62	156.30	156.79
72%	73%	74%	75%	76%	77%	78%	79%	80%
157.28	157.77	158.00	158.00	158.24	158.73	159.00	159.00	159.00
81%	82%	83%	84%	85%	86%	87%	88%	89%
159.00	159.36	160.34	161.32	162.30	163.14	163.63	164.00	164.00
90%	91%	92%	93%	94%	95%	96%	97%	98%
164.00	164.00	164.08	164.57	165.00	165.00	165.04	165.53	166.00
99%	100%							
166.00	166.00							

And similarly, in case if you try to make here 100 partitions, that means you have to control only the cross function by writing sequence 0, 1, 0.01 and it will give you here such 100 values you can see here, this is here, the 100 values and hundredth value actually, so, this is our percentile.

And similarly, if you want to make it here at an interval of say here 0.33, so even that is possible, you can see here this is giving you any type of value. So, remember one thing we are going to use this function, when we are trying to use the part, when we are going to learn the, the part of a statistical inference at that moment, I will remind you that we already have computed or we already have learnt how to compute these different types of quantiles or percentiles.

(Refer Slide Time: 33:47)

Tschebyschev's Inequality:
If we do not know the distribution of a random variable X , we can still make statements about the probability using Tschebyschev's inequality that X takes values in a certain interval (which has to be symmetric around the expectation μ) if the mean μ and the variance σ^2 of X are known.

21

Now, I come to the last topic of this lecture, which is about the Tschebyschev inequality. So one thing you have to be very careful that how to say this word. So this is Tschebyschev. So now, this Tschebyschev inequality helps us in those situation, when we do not know the distribution of the random variable X , but we still want to make some statement about the probability?

So, even when we do not have any idea about the probability distribution of X , well at this moment, you may not understand but just after this, I am going to start with different types of probability distribution functions and probability mass function. So, you will see that the random variable will always have some probability function and now suppose you do not know, but if you want to make certain statement about the probability, then we can use the concept of Tschebyschev inequality.

And in this case, the X takes values in a certain interval and the condition is this this value has to be symmetric around the expectation that is mean μ . So, in case if X is taking the values in certain interval, if the mean is μ and the variance of X is σ^2 and both μ and σ^2 are known to us.

(Refer Slide Time: 35:28)

Tschebyshev's Inequality:
Let X be a random variable with $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$. It holds that

$$P(|X - \mu| \geq c) \leq \frac{\text{Var}(X)}{c^2}$$

x: pdf pmt ?

This is equivalent to

$$P(|X - \mu| < c) \leq 1 - \frac{\text{Var}(X)}{c^2}$$

|x - μ| < c → Deviation from mean
-c < x - μ < c
μ - c < x < μ + c

In that case, the statement of Tschebyshev inequality is that, that if expected value of X is μ and variance of X is σ^2 , remember one thing what will be the pdf or pmf of X that we do not know in this case, the statement is simply trying to say here $P(|X - \mu| \geq c) \leq \frac{\text{Var}(X)}{c^2}$.

And same thing can also be written in this format at the $P(|X - \mu| < c) \geq 1 - \frac{\text{Var}(X)}{c^2}$. Now, if you try to see sometimes students get badly confused with the symbol and notation, but this is very simple, what are you trying to say, if you are trying to say here $x - \mu$ is a suppose less than c What does this mean? So, this means $x - \mu$ is going to lie between $-c$ and $+c$ that means, x is going to lie between $\mu - c$ and $\mu + c$.

So, if you try to see here, when you are trying to take here this quantity $X - \mu$. This is only the deviation from mean and this quantity is very important that whenever you get your marks in a test, you always try to measure it from your mean well, whenever you go to your home and your parents asked you that, how is your performance in the exam, you always try to compare your marks with the class average in case if the difference between your marks and class average is quite large, then possibly you are on the extreme.

For example, if the class average is suppose 60 marks and you have got only say 20 marks, then you are, then the performance in the examination is bad and if you have got say 90 percent marks that means the performance is very good in the examination, but in case if the class average is suppose 90 percent and somebody has got the mark say just a 91 percent or

70 percent, this will have a different interpretation. So, we want to compute this type of probability.

(Refer Slide Time: 38:02)

Tschebyschev's Inequality: Example

Consider the continuous random variable "waiting time for the train". Suppose that a train arrives every 20 min. Therefore, the waiting time of a particular person is random and can be any time contained in the interval $[0, 20]$. The required probability density function is

$$f(x) = \begin{cases} \frac{1}{20} & \text{for } 0 \leq x \leq 20 \\ 0 & \text{otherwise.} \end{cases}$$

$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^{20} x \frac{1}{20} dx = 10$

$\sigma^2 = Var(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x)dx = \int_0^{20} (x - 10)^2 \frac{1}{20} dx = \frac{100}{3}$

We can calculate the probability of waiting between $10 - 7 = 3$ and $10 + 7 = 17$ min:

So, let me try to take some small example to show you that what will be the difference between computation of a probability directly from the probability function and using this Tschebyschev inequality. So, let me try to take the same example that I have considered earlier many times that there is a continuous random variable indicating the waiting time of the train and suppose the train arrives after every 20 minutes. So, the waiting time for a particular person will be random between 0 and 20 minutes and required probability density function in this case, this is here $f(x)$ is equal to $1/20$.

Now, suppose we do not know this $f(x)$, but suppose the value of here mean and variance and for that mean, so we can just compute it. So, mean is going to be that we already have computed that expected value of x which is coming out to be 10 and σ^2 is the variance of X which is coming out to be $100/3$ that we already had computed in the earlier lecture.

So, now suppose you do not know the PDF, but you simply know the value of mean and variance that is μ and σ^2 . So, now we can collect calculate the probability of waiting time between $10 - 7 = 3$ and $10 + 7 = 17$ minutes.

(Refer Slide Time: 39:21)

Tschebyschev's Inequality: Example

$$P(|X - \mu| < c) \leq 1 - \frac{\text{Var}(X)}{c^2}$$
$$P(|X - 10| < 7) \leq 1 - \frac{100/3}{7^2} \cong 0.32$$

The probability is therefore at least 0.32.

However, if we apply our distributional knowledge that

$$F(x) = \begin{cases} \frac{x}{20} & \text{for } 0 \leq x \leq 20 \\ 0 & \text{otherwise,} \end{cases}$$

then we obtain a much more precise result which is

$$P(3 < X < 17) = F(17) - F(3) = \frac{17}{20} - \frac{3}{20} = 0.7.$$

So, now using the Tschebyschev inequality I can write down here, probability that $|X - \mu|$ is smaller than c that can be written as probability that $|X - 10 - 10|$, μ is here 10 is less than 7 is that deviation from the say, $X - \mu$ you can see here in this question, this is what we are writing here. Now, this is going to be less than or equal to $1 - \text{variance of } X / \text{variance of } X$ that we have computed to be 100 divided by 3 and this is c square. Square is here 7 square, because c here is 7 and if you try to approximate it this will come out to be 0.32.

So, probability is therefore at least 32 percent that is what you have to understand that is the interpretation that the probability is therefore at least. Now, in case if you try to say use this probability function and we try to obtain the distribution function that will come out to be like this capital $F(x)$ equal to $x/24x$ line between 0 to 20 and if we try to obtain the more precise result for this event, then I can simply write down here that X is lying between 3 and 17 which is equal to $F(17) - F(3)$ that we already have learned that if you want to find out search probabilities using the CDF how you can find it out.

So, this will come out to be here $17/20$ upon $- 3/20$ that is the same example that I had conducted that I had considered earlier also. So, this probability comes out to here 0.7. So, now you can see here, this is the exact probability and this is here the probability in terms of at least.

(Refer Slide Time: 41:15)

Tschebyschev's Inequality: Example

$$P(|X - 10| < 7) \leq 1 - \frac{100/3}{7^2} \cong 0.32 \quad \checkmark$$

The probability is therefore at least 0.32.

$$P(3 < X < 17) = F(17) - F(3) = \frac{17}{20} - \frac{3}{20} = 0.7. \quad \checkmark$$

We can clearly see that Tschebyschev's inequality gives us the correct answer, that is $P(3 < X < 17)$ is greater 0.32.

The exact probability is 0.7, is rather poor for approximate probability.

One needs to keep in mind that only the lack of distributional knowledge makes the inequality useful.

25

So, what you can see here that connection from the Tschebyschev inequality you are getting the probability at least which is 0.3 to approximately and for and using the pdf you are getting this probability to be 0.7. So, we can very clearly see here that the Tschebyschev inequality gives us the correct answer that is probability that X is lying between 3 and 70 is greater than 0.32. The exact probability here is 0.7, is rather, is rather poor for approximate probability. Well, it depends that what type of c, are you going to choose?

So, my idea is that I am not trying to criticise the Tschebyschev inequality or any probability, but I want to explain you here both the concept and depending on the situation you have to choose what you want to do whether you want to employ the Tschebyschev inequality or you want to compute the exact probability that depends on your objective that I cannot tell you at this moment, my objective is to make you learn all the possible thing.

So, one needs to keep in mind that only the lack of distributional knowledge makes the inequality useful, but because once you do not know that what is your probability function, what are you going to do you have to choose something, so in that way, this is going to help you.

So, now, we come to an end to this lecture and you can see here we have consider, the topics of moments quantiles in this chapter in this section and the moments are going to give us a very realistic information about what is happening inside the data set and then quantiles are also giving us the values of the partitioning the data set. So, these are the different partitions.

Now, it depends on you in data science, whenever you are trying to deal with a huge data set you cannot view those data set by your eyes, there can be million, billions and trillions of observations. Say not only we gigabyte or but that terabyte, the data may be in terms of petabytes. So, you cannot view that data with your eyes, but you have to depend on these tools to take out the hidden information from the data.

This hidden information will be the statistical information and the statistical information will be in terms of moments like as mean variance skewness, kurtosis, etc. Now, the question is where are you going to use mean and where you are going to use your skewness that is up to you, but my opinion and my suggestion is that when the data comes to you the data cannot display data cannot raise the hand and can say, I have this this property. So, we tried to implement all sorts of tool graphical analytical, we will try to find out the mean variance, skewness, kurtosis and finally, we try to get gather different type of information so that we can take a correct conclusion.

So, now, I will say try to take some example from the assignment from the books and try to understand these topics, they are surely going to help you when you are really going to work in data science. So, you try to practice and I will see you in the next lecture with a new topic on probability functions. Till then, goodbye.