Science Communication, Research Productivity and Data Analytics using Open Source Software

Dr. Mohit Garg

Central Library

IIT Delhi

Week: 06

Lecture 22: Descriptive Analysis II

Hello Learners, welcome again. So, in the last lecture I have discussed how we can import the data in R. Then we have done some of the basic computation like frequency analysis. And under frequency analysis we have calculated the simple frequency. then we have calculated the cumulative frequency and after that we have seen how we can do the relative frequency on a particular dataset. We have also seen it in both ways. we have seen through manually then we have seen in the R and in the R we have seen in a lengthy way also. then in the past we have seen it in shorter ways with just two commands. So, before discussing the other descriptive analysis concept like measure of central tendency, let us first discuss functions. And if you have noticed that from week 4 and in this week also I am talking about a lot many functions. So, what are these functions? So, let us understand this concept of function and then we will go ahead to the measure of central tendency and then we will also create our own function for our different tasks. So, you can consider a function as a kind of a switch that is written for doing one kind of analysis.

So, what exactly is it? So, if you have seen this mean() function. Then this mean() function is used for calculating the average of the values. Then we have also seen sum(). So, this sum() is to calculate the sum of the values. Then we have seen the dim() function, we have seen nrow(), we have seen ncol(), we have seen data.frame(), we have seen c(), we have seen even seq() in the last lecture. This is a shortcut to this function sequence. So, there are many functions there and this function made our task easier. So, each time we would not be doing this kind of long analysis. If we will be doing this way it will take our huge time.

So, that is why R has a lot many functions and it has some built-in functions and some functions we can use by using the packages. So, first we will create our own functions. So, to create a function in R you have to remember this particular syntax. So, the syntax to create a function in R is you have to name a function (namefunct). So, this is my name of the function and then you have to assign it and then this is my function.

So, you have to call it like exactly what it is saying. So, this is a function. So, this assignment is saying that this is a function and what exactly it will do? It will do exactly

what is mentioned here. So, whatever is given here it will do the task. So, this is a one time task. And Once you have done this task, you do not have to do it again and again.

It will be saved. We will see how we will create these functions in R. So, here I am giving my function as a name Mohit. And I am calling this function and then I am assigning that you have to do this task. So, what function am I creating? I am creating a function named it as a Mohit and it will give the result every time whenever we call it. That is welcome to the NPTEL course. So, if I do like this. Now if I run this function. So, as soon as this function is called it will retrieve this value only. Like say for example I am doing like this. So, the task assigned to this function is that you have to print "Welcome to the NPTEL Course". That is so this is very simple. So, this is how we can create a function.

Let us now create one more function. So, I am creating another function named Mukesh and using this. So, this is my name for the function. And I am saying that this is a function. Now I am saying that the task of this function Mukesh is to print "Hi everyone! I am a teaching assistant for this course". So, if I run this and now if I call this function. And now if I call it Mohit, welcome to the NPTEL course. So, we have created the two functions whose task is to just print the texts. There are no operations there. The task is simply just whenever they are called they have to print this string.

Now, if you have seen here that in this mean function or in the sum function or in this dim function they require something to insert. So, if we run this sum function like say for example if I simply run this function sum 0, because there is nothing present or if I do like this it means something is missing. So, all these functions require something to do the computation. Once you assign the value, once you give it here like 1, 2, 3, 4, 5 or here, it will calculate the mean or maybe the sum and it will give the value. But the function that we have created for Mukesh, Mohit for these two functions there is no such requirements. So, the only thing is that whenever this function is called it will give the results.

So, whenever we will be creating a function that will give some argument. then only our task will be easier for our data analysis. So, we will now create a function, which is argument based. So, this particular value whatever comes here is known as an argument. Now we will create one function with a single argument and then we will create the function with double arguments. So, say for example I am creating a function 'Priyanka'. And what I am doing is whatever the value is assigned here it will be multiplied by 2. If I am running this now, if I call Priyanka, see the same error. Whatever the error came from, it is like an argument is missing in me. The same error is showing here. So, what my value is say for example now I am taking 3. So, now I am taking 3. If I run like this. So, what exactly happens as soon as the value is assigned here to there is 3. So, the function takes that value and multiplies it by 2 and gives the results.

Or like say for example I am doing like this 345 and now first I am creating this object. And now if I am calling this way it is showing 690. If you have noticed here, that is exactly what I have done. So, writing like this, or like this is also giving the results. It is also giving the results.

So, though it is working but it is a standard practice do not give space after your like name of the function. There is no such requirement of giving the space. So, always give the name of the function and then the brackets and then your argument whatever the argument is there in that function. Now if you see here that we have created 3 functions one is Mohit, one is Mukesh and another is Priyanka. All these 3 functions I have just taken as example one. But they are like they are not, these names are not saying what task they are doing.

If you see the example here that this mean function is there. So, we can understand that this particular function calculates the mean. or this particular function calculates the sum. But instead of this if we use the name Mukesh. We can create a function called Mukesh with sum, which does the addition of the values. We can create a function named as Mohit or some other names which do the whole operation. But whenever you will be having a lot of functions like 10 to 20, for practice or for understanding the concept of function it is like recommended that you create this kind of names of a function. But whenever you are creating a function for your analysis for a long term always use the specific name functions.

So, say for example you are creating a function to calculate the square root. So, instead of keeping the name Mukesh you can use this function like square_r or square_root like that. So, it will be easier for you to create the function on the same day and doing the analysis on that particular day will be easier to know. I have created a function called Mohit which will result in the string Welcome to NPTEL course. But like say for example I am doing after sometime maybe after one year I may not remember what exactly I have assigned. So, it is a good practice that you give some names that are very intuitive that give the idea of what exactly this function will do.

So, until now we have discussed two kinds of functions. So, one kind of a function is Mohit and Mukesh where no argument is required. And then there is a second kind of a function where an argument is required like Priyanka. To run the function Priyanka you have to give some arguments. Now we will see the example of two arguments. if we have to create a function with two arguments. So, I am creating a function called AddNum.

So, what AddNum is doing? AddNum is adding the two numbers and it requires two arguments and whenever the two values are given the addition to be written. If I run this, if I run this so I get this my function is created. Now if I do like this. So, this is my function that adds numbers. Now if I do like these unused arguments.

So, this is the error. It will come because there are only two arguments there. But for some functions if you do like that it is working fine. And it will also work fine with two values also. So, this is based on how the function has been written and what arguments are given and what are the standard requirements. So, that is what you have to remember whenever you are creating a function. If you are creating a function you have the given the proper kind of arguments. You can't use a function without an argument and then you will calculate any of the operations. Now like say for example if I am doing like this Mukesh(2,3). Again it is showing here is the unused argument. So, this unused argument is showing that it doesn't require something to be missing. Either the argument is not defined or like it is a function where no argument is required. For this Mukesh function no argument is required. If I do like this and if I run this Mukesh my function is already created.

So, it does not require arguments. Now take it in another example where we are calling multiple operations and then we will see what exactly the result is. I have created this function and now if I run this. It is 5. So, how come these 5 come? This 5 came from here 2+3. If I do here like this b*3. Now if I run this, it is 9. Because a is here 2 and b is 3. And our last operation is b*3. So, you have to remember that whenever you are defining the function it will return the last operation value. Now take another example if you want to use the earlier operations. So, what happened here is that here these were computed but these values were not assigned to a. That is why a plus b was 2+3, 5. So, we will get the answer like 2*2, 4 and 4+3, 7. So, to do that what we have to do is we have to assign the value first that whatever the operation is there to a and then we have to get the result. So, if I do like this, now if I run this my result is 7. Earlier my result was, if you see here, my result was 5. Because if we see here only the value 5 came. And this value 5 came because only this operation was computed by the function not this operation.

But here what exactly happened to this particular a*2. Whatever the value is there it is assigned to a. And then on this a the next operation is done a+b and that is how we got the value 7. So, this is how you can create your own functions and you can write your task specific function whatever you want to do the operations. So, as a practice I request you to create some functionS with your name with different kinds of operations and share them with us in the Discussion Forum. Now I will discuss the measure of central tendency and how we can calculate this mean median mode in R.

So, to start with the first thing like we import the data we have already imported let again first see our dataset. So, this is our dataset if I run this. So, this is our dataset where values are there. So, in 2019 there is one publication, in 2021 there is one, 2022 there are 5 publications 2023 one publication. Now how we used to calculate the mean manually. So, to calculate the mean of what we used to do. we do the sum of these values. So, this value sum is 8 and then the number of values n.

So, n is here 4. So, the mean is 2. Very simple. the same way we use the mean in R. So, to calculate the mean in R first we will do the sum of those frequencies. So, how will we do that? So, we will use the sum function as soon as we run this. Our 8 came what the value came here at 8. then I am assigned to this. Now if I run, this is done. Now we will calculate the number of observations. So, how we will do that we will use this function length(). So, this is 4. Now if I run this and if I check the value of this 4. So, we got the sum of our frequency: it is 8. And we have got the number of observations that is 4. So, now what I mean is the ratio of the sum of the frequencies by the number of observations.

So, I will do it like this. So, this is my sum and this is my number of observations. If I run like this, then it is 2. So, what are the results we have got from our manual analysis? The same answer we got here. So, this means Fq returns the 2 values. So, this is how we have calculated the mean. So, by default there is like one default function is there in R which is used to calculate the mean. But if we do not have any such function, we can create our own function also.

How will we do that? What we have to do is to just simply call this function Fq1() . It is a function and so it takes only one value, that is frequency. Now after that what operations are we doing? We are first doing the sum (Fq) value. So, whatever the value is assigned we are first seeing the sum and after that we are calculating the length. then we are doing like this. So, this is how we can create our own function, like here what we have created means Fq1 is a function which is used to calculate the mean in a dataset.

We do not want to use the inbuilt function. So, this is how it gives a better understanding of how the concept works in R. How we can define our own functions. So, we can save this. And the next time whenever we have to use this mean Fq1 we can just simply run it and we can call it this function. So, now though in this case we have an inbuilt function in R that is mean.

So, we will check the result with that also. If I do like this, the answer is 2. So, this answer is also 2. And the manual we have calculated is also 2. So, the average value of publications is 2. So, this is how you can use the functions in R by either creating your own functions or using the inbuilt function or by using some external packages and then importing those packages. Now, we will discuss the second concept in the measure of central tendency that is median.

So, median is the kind of a measure that is not affected by the like outliers, or if there are any extreme values. So, say for example if you see in this particular dataset. This 5 is an extreme value. If you see this kind of a dataset. There is a common average kind of thing there. like if we remove this 5 from this dataset, and then we will calculate the average. So, in 2019, 2021, 2023 the average publication was 1.

But in 2022 there are like 5 publications. This is one outlier, a kind of an outlier which affected the mean. So, that is why we use median to calculate central measure exactly where the data is. So, to calculate the median manually is what we used to do. First we arrange the data in either ascending or descending order. So, like say for example this, this, this, this. Now, if the number of observations are odd then it will be whatever the length will be there it will be the middle value like if it is like one more value is there.

So, the median is these 2 values. So, in our case the number of values are even and these are 4 values. So, what median is median is the second value and third value 1+1/2 that is 1. If we see the mean it was 2 and our median is 1. How we will calculate this. Let us see now. So, to calculate the median what we have to do is first we have to order the data.

So, we will be ordering the data. our data is ordered. So, what exactly does it return? It returns that this value is at 1, 2 and this particular value because our dataset if you see here is DfdatFq 1,1,5,1. So, what exactly it is showing here is that. this will be at the first location, this will be at the second location, but this 5 will go at the fourth location and this will come at the third location. So, that is done. And now we will do the order ok. So, our ordering is done in this. Because that was the highest value 5. And we have kept it as the fourth location. Now we will be counting the number of observations that whether we have even or odd here the number of observations are even. So, the median will be the average of the two middle values in our case the number of observations are 4. So, what we have to do is to calculate the average. The average will be an ordered dataset. And at the second value, and then plus, the set 2 and the third value, and divided by 2. If I run this our median is 1 and we can now say medDf. So, our dataset has a median of 1.

So, that value we have allocated to medDf. So, this is how we can calculate the median by our own calculation. But there is one inbuilt function. But we can use it to calculate the median. we will see now. So, this median is a function which is used to calculate the median. So, now if I run this is 1. So, what we have done is as a lengthy way. We have first ordered the dataset. And then we have seen what the number of values are there and like in our case it was even. So, we have calculated the average of the second and third value. And then after that we have assigned that value to a particular object that is a median value.

But here this inbuilt function is there for the median you do not have to do all these operations. But this is how you will get the understanding of how these functions are working. So, our mean is done, median is done. So, now in R there are no inbuilt functions available for calculating the mode. We have the inbuilt functions for mean. we have the functions for median. But we do not have the functions for mode. So, how to do that? So, again we have to write a function, but to write a function we must first understand how we will calculate the mode in a manual way.

How we used to calculate when we have data using pen and paper. So, first we will calculate the mode manually and then we will see how we can create a mode in R by creating our own function. Or packages are also available which we can use to calculate the mode. I will show you that also. So, mode is a measure of a central tendency which is used for counting the number of occurrences of categorical data. Say for example in 2019 someone has published five publications. And out of these five publications three are in open access. And the other two are in closed access.

Now here the mode is the most occurrence. So, the most occurrence value is open access. Because the occurrence of open access is 3 and closed access is 2. So, this is how we calculate the mode as a manual way. Now we will see how we will calculate the mode in R. So, to calculate the mode, the first thing we have to do is to count the occurrence.

So, here the most occurring value is like 1. In this case, if I see here for this particular case I am assigning it because this name is becoming too big. I am assigning to another object called x. So, this x is just a frequency. So, I will just table it. So, what exactly, so this was our dataset 1, 1, 5, 1; 2019,1, 2021,1, 2022, 5, and 2023, 1.

Now we have tabled it. Tabled it means we have tabled these 1,1,1. So, the occurrence of 1 is 3 times and the occurrence of 5 is 1. So, this is what it is showing. So, I will create this table. Now we have already calculated the occurrence of each of the values. So, there are like 1 that occurs 3 times and 5 that occurs only once. So, we have to see which value occurred the maximum time. So, for that we will use the maximum function and if you see that it is showing that 3. So, the 3 is the maximum time a value has appeared. So, I will assign it to this and this is our value.

So, there is a value 1 which appears 3 times. Now we will be checking the value, which occurred the maximum time. So, in our table, we will be using that again. So, if I do like this. it will check where exactly the number of frequencies is highest. So, for the occurrence of 1, it is showing true that means the value 3 is for this 1. And this is showing false. Because 5 doesn't appear 3 times. So, that is why it is showing false. Now we will only select the value which has appeared the maximum time. So, this is it. So, that is our 1. Now if I give the names to this, this is 1.

So, 1 is like if you have seen here like this 1 appears in the string. It is not numeric and our data is a kind of numeric. But for this understanding I considered it as categorical data. So, I will make it numeric and then like I will assign it to modeValue(). Now if I check the modeValue(), this is 1 and the maximum frequency is 3. So, in this particular data I am not talking about the 2019, 2020 the whole earlier dataset. or I am talking about this data where we have occurrences of 1, 3 times and 5 is only 1 time. That is why the mode of this dataset is the value 1 which has occurred 3 times.

Now what are the things we have done step by step? We will create a function the same way we have created the function for mean, here mean Fq1. in the same way we will create the function for mode. So, all these operations will put it under the function time. So, this is created. Now if I do like this, it is showing that there is a value which has occurred 3 times.

If you have noticed here that only 1 value is printed. 3 is printed, not 1. So, this was what I was discussing when we were creating the function. So, whatever the last operation is, it gives the value of that only. But there is a better way to calculate the mode of a dataset than using a package. We have already discussed how to install the package. So, to calculate the mode in R we need a package called "DescTools". And we can install it here also. And it is showing here.

So, I think we will use this option and this package will be installed. After the installation of this package we have to call the library. We will call this library and now we will run the mode function. We have defined our function as nMode. But they have defined the function as mode. So, I will run this. So, it is shown here that that 1 is the value which has occurred 3 times. So, this is about the 3 of the measures of central tendency that one is mean, median, mode. We have discussed how we can create our own functions and there are some inbuilt functions that are there. And also how we can use the functions which are in some external package. So, for mode we have created our own function also and we have used the function of the external package. So, I request you all to use the practice dataset and perform all these operations, calculate the mean, median, mode by calculating the lengthy way, and then in the shorter way.

Let us know your experience of doing the analysis. I will see you in the next lecture where I will discuss the measure of dispersion. Thank you.