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

Dr. Mohit Garg

Central Library

IIT Delhi

Week: 06

Lecture 21: Descriptive Analysis I

Done

Hello Learners, until now we have discussed science communication, different channels of communicating science. Then we discussed the different research metrics. After that we have discussed the different sources of extraction of data, like we have discussed about Scopus, then we have discussed about Web of Science and PubMed and how we can extract the data from these particular databases. After that we have also seen about the basics of R, how we can install the R and Rstudio and what the different objects and different object types are. And after that we have also seen the three of the theoretical laws like Bradford's law that helps in identifying the core sources. Then we have also seen the Lotka's law which tells about the scientific productivity of the authors and then in the last we have also seen the Zipf's law which is related to the frequency of words.

So, in this particular week we will be discussing descriptive analysis. Descriptive analysis helps in understanding what given data is. So, it summarizes the dataset to identify the patterns and trends and relationships. So, we get insights into what has been done without making predictions or prescribing any action.

So, in descriptive analytics we are not predicting anything, just we are explaining that this is the thing. This has been done. So, frequency is the core component of descriptive analytics and basically where we count the occurrence of a value in a dataset. Frequency analysis gives the answers to questions like how common are certain values, what is the distribution of values across the dataset, are there any outliers of patterns that we need to know. So, if we talk about in terms of research productivity or related to this course, frequency analysis can derive the answers like who are the top prolific authors in a particular journal or who are the prolific authors in a particular country. So, this kind of analysis can be done with the help of descriptive analytics.

So, in order to do the analysis on scientific productivity we need to handle the two key parameters. So, one key parameter is the publication and another is citation and we need to do counting of these two parameters whenever we will be doing the assessment of

research productivity. But during the counting we need to take care of a few of the things. So, the first and very important thing is the data source we are using. The selection of the data source is very critical as there are different data sources available like Scopus, Web of Science, PubMed, etc. and all these differ from each other in terms of coverage and scope. So, the next thing we have to take care of is the types of documents we are considering for the analysis. Like we have already discussed , there are different types of documents published like the conference papers, book chapters, research articles, review papers, notes, then data papers, editorials, etc. It is again an important thing to note down what document types you will be considering for assessing the unit of a publication.

And the last is related to citation. In citation we can have self-citation, local citation, global citation or maybe what kind of sources we have selected for receiving the citations. So, let us understand all these points with some examples then you will have more clarity on why I am saying that these points should be considered before the assessment of the research productivity. So, say for example this is Scientist A, and this is Scientist B. Now, Scientist A has 8 publications and 5 citations and Scientist B has let us say 10 publications and 3 citations. So, in this kind of a situation we can easily say that Scientist B has more publications compared to Scientist A, and we can say that Scientist B is more productive than Scientist A.

And based on this assessment we can easily assign the reward to Scientist B. So, now take the second example where we will see this publication in a deeper sense. So, what exactly is it? So, say for example we are looking for a particular kind of data source. So, I am taking Scopus as an example. So, now I am considering Scopus as a database for the analysis.

So, now we will say that out of 8 publications Scientist A has 5 publications in Scopus. but Scientist B has only 3 publications in Scopus. If we see this scenario, Scientist A has 5 publications in Scopus and Scientist B has 3 publications in Scopus. If we see only this scenario, forget about the earlier case. If we see only this through Scopus then we will say that Scientist A is more productive than Scientist B. And now we will say that Scientist A is the winner.

Now take another example of the kinds of document types. So, if we talk about the number of documents, this Scientist A has published 2 publications as research articles and this Scientist B has published all 3 publications as research articles. Now if you are doing based on the analysis that you are considering the Scopus database. then if you are assessing only the research article, other documents like the ranking agency used to do. So, the ranking agency does not consider all kinds of document types. So, if you are focusing only on the research article. then in this case Scientist A has 2 publications and Scientist B has 3 publications as a research article in Scopus. Now in this case Scientist B will be the winner. So, in the first example, who was the winner? B was the winner. In the

second example, who was the winner? A was the winner. Or in the third example, who is the winner? Again B.

So, this is why I am saying that it is very important what kind of data source you are taking in. and what kind of documents you are considering for the analysis of a research productivity of a particular author or particular country or particular institution.

So, similarly the case of citation can also be seen, like the author has 5 citations and this Scientist B has 3 citations. If I say that out of these 5 citations Scientist A has 1 citation from any international source and other 4 citations are from self. But now if I talk about Scientist B it has all 3 citations from international sources. So, the concept of citations, self-citation and things like these global and local issues are still in gray. So, we cannot say that exactly this self-citation is in black and another kind of citation is in the white.

But it is always worth considering these points whenever we count the publication and the citations. So, once these points are considered let us now start the descriptive analytics. So, it includes a measure of central tendency and a measure of dispersion. We will see each one by one by analyzing the data manually and then we will see in the R. And I am doing the manual analysis of a few of the concepts of descriptive analytics to give you more understanding about that particular concept, and how you have to use the R for calculating those things.

So, let us start the first thing. So, let us consider that this is data extracted from Scopus with only 3 variables. Title of the paper, year published and then number of citations it has received. So, this is the title, then year, and cited by. And by default each download in Scopus comes with this extra column of a link. We will delete this and consider the only 3 for our analysis. So, we will not consider this. and we will take care of only these 3 variables. So, our dataset is this. So, the first thing is what we used to do is count the number of publications. So, here what we will be doing is count the number of publications year-wise.

So, if we see here like in this particular dataset, the number of unique values. So, we have 2023, then we have 2022, then we have 2021, then we have 2019. Now we will count the number of publications in each year. So, in 2023 the occurrence of 2023 is only 1. So, this is 1, then 2022, 5, then 2021, 1 and then 2019, 1.

So, this is our frequency of this particular dataset. Now we can easily infer that for this particular kind of a dataset there are like 5 numbers of documents in the year 2022. So, from here the result shows that in the year 2022 the author or the particular institution has the highest number of publications, that is 5. Okay, now we will do the same thing. So, if you notice this particular thing, what it's saying is that the package, desktop and moments are required but are not installed. Okay. So, if you use this script I will be sharing the particular script on the github. So, it will show because the package is not installed. Okay.

So, again this is the beauty of RStudio that it automatically pops up that this package is required if you install it. Okay, we will see later in the part where this package is required.. Okay, we have already extracted the data.

Now the very first thing we will be doing is to import the data into RStudio. So, here we will be importing the data. So, to import the data there is a function called read.csv because our dataset is in the CSV file. Okay. So, Scopus provides the features of downloading the data in different formats.

So, I have downloaded this particular kind of data in CSV format. So, I am running this function read.csv. but it has many other functions that are also available for importing the other kind of dataset. Okay. So, whatever the kind of dataset you have you can import into R. So, I am assuming that you have the read.csv. but default many times we are used to downloading CSV or Excel files from the data sources. So, we will be using the read.csv function to import the data. So, this is the read.csv and this is my dataset. So, what I will do. I will just run this command and if I run this my data is imported.

So, after importing the dataset let us now see what exactly the data is. So, if I see this data. Okay. So, this is the dataset we see here. if we see here this is our dataset. So, we have imported this into R studio and we have assigned this to a name that is Dat.. Okay, you can assign the name but like what I have discussed in the previous weeks.

Okay that it depends totally on you what names you give. So, I am giving the name Dat. Again an important thing you must note it down here that I have just mentioned the only file name of my dataset because my current directory is in the documents.. Okay, and where this data is available.. Okay, if your data is available in another directory then you must change either the directory first or you give the proper full path. Okay. So, say for example if your data is in the download directory.

Okay. So, you have to give that path like this. Okay. So, I have commented below. So, this is my download directory and this is where my data is. Okay, but if you see my current directory is documents and this file is also in the document directory. So, this is where my file is. So, that is why I have just given only the file name. if your data is in the download directory and if you haven't changed then you have to give the proper path. and if you change your directory by using the functions set working directory to downloads then only you have to name the data. Okay. So, as a good practice it is always recommended that whenever you are starting the analysis, or whenever you are importing the data you must check what exactly your current directory is.

Okay after this like our data is imported, and we have also seen that data is correctly imported. So, here we can see that data is correctly imported. Now as per our counting of the publications we were doing manually. We counted what were the unique values in the dataset. So, those were the years. So, those years were 2023, 2022, 2021, 2019. Okay. So,

the same thing we will do here also. So, our data is Dis and it has three columns. Okay, and we will access that particular column only. So, we are not doing the analysis on title or the cited by we are only counting the year-wise that number of documents. So, for that we will be considering that particular column here also we have considered this particular column year only. So, I will do this here. So, unique here if I run this. So, it is saying that these are the four values which are unique 2023, 2022, 2021, 2019. and that was what we have seen in the manual also. Okay. So, for this calculating the unique values in the dataset we have used unique concerns.

Now we will assign this value to a particular object. So, we have assigned it to this. So, if we see the value of this. Okay. So, now the next thing after finding the unique values in a dataset. So, we need to know the occurrence of that particular value. Okay. So, for that we will be using the function table. So, what table does. The table just counts that unique value and also counts the frequency of those unique values. Okay. So, if we run this. So, now this table function has given these values that in 2019 there is only one publication. 2021 this is one, 2022 there are five publications, 2023 there is one publication. Okay. So, what table is giving a table is giving that there are like five occurrences of 2022, and all the other threes are only single occurrences.

Okay and then based on this analysis we can infer that 2022 was the most productive year for this particular dataset. Okay. So, if you have noticed here that the earlier output of this function was there is no where this where one and frequency is there. Okay. So, as soon as we do the data framing of this particular table this column name is assigned. Okay. So, I am assigning this to an object now. if we see here that there are two column names where one and frequency these may create some like whenever we are doing the analysis may create some confusion. So, for better we will be changing these column names. So, if you remember, we have already discussed how to change the column names. So, we will change the column names like this. So, I am changing the column name of where one to here and the free to frequency. So, if I run this and now if I will see the value of this. So, now our one analysis is done but we have done it manually. Okay. Now we will see this dataset. So, this is our dataset, and we will save it as an external file. So, like the way we have imported the dataset in a similar way we can export the data also.

Okay. So, to export the data whatever we are working on and whatever our analysis is we will be using the function called write.csv. Okay, and these function names are based on what task they are doing. Okay. So, if you see the function name read.csv it means that it's reading the data it's not doing anything but if you see the name of another function which we are using for exporting the data it is write.csv. So, it is writing the dataset. So, I am using this write.csv. This is my particular dataset and I am giving the name of that file. Okay, write.csv. So, this file is created. So, again here also path matters. Okay, if your current directory is documents it will be saved in the current directory only in

documents. but if your current directory is downloads or desktop it will be saved in desktop or downloads.

Okay. So, we'll see here documents and if I just refresh. So, frequency analysis is there. So, one analysis is done. So, now we have done the analysis first and then we have also saved the analysis to our external file now we will move to the second part of cumulative frequency. Okay. So, to calculate the cumulative frequency manually is what we used to do. So, for 2019, 2021, 2022, 2023 this is our dataset. In 2019 there is one publication this is one there is five and again this is one. Okay. So, this is our frequency now. What will be like this will be one then this will be one plus one and this will be one plus five plus one, and this will be five plus one plus one plus one. So, this is one, this is two, this is seven and this is eight. So, now we will see in R how we can calculate the cumulative frequency.

So, now we don't know in R how to calculate the cumulative frequency. Okay, how we will do that. We have already discussed in week four that we will seek help. and how we will seek the help we'll see by using the fourth quadrant. So, we'll go here and we'll write a cumulative and if I enter this many results will come and what I am looking for. I am looking for this cumulative sum if I click here. So, it is showing the example. Here are the examples of how I have to calculate the cumulative sum. So, if I use this function. So, what exactly is this argument? So, it generated the sequence of ten values. So, arithmetic progression with the common difference of one. So, if I do like this here one to ten there are ten values there. and the cumulative sum of these ten values is like this is one then this is two plus one then this is three plus two plus one. This is how the cumulative sum is generated. So, the same thing we will be doing on our dataset. So, our dataset is this. So, what we will be doing is we will be giving our dataset and then we will hold this one son. So, now if I run this. So, we have got our cumulative frequency. Okay. So, now we will convert this to a data frame. So, I have converted to a data frame. So, this is my data frame. So, this is a year and this is a cumulative data frequency. So, I'll run this and our new data frame is created. Okay. now if you see here that this column name is too big Okay. and what exactly how this column name comes in. So, this is saying that this is a cumulative sum of this particular dataset. Okay. So, that is why some dot FQ is there. Okay. So, this is also a kind of an identifier that this particular column is related to a cumulative sum of the frequency. by default the data frame takes the value whatever it is assigned. So, that is why it takes this value as a column name. So, now we will combine this cumulative frequency with our earlier analysis. So, to do that we have to create a data frame. We already have one data frame which we have here and this frequency now we will add the cumulative frequency. So, how will we do that? So, this we already have and we will add one more column from this particular data frame. So, this is our data frame and I have used dollar signs. So, if I just remove this ,and if I hit that tab it will automatically come. Okay. now if I run this. So, my data frame is created. So, this is my

year, then the frequency, then the cumulative frequency. Okay. Now if you see this. This column name is too big. Okay. like we have to make it shorter column name for like. because when we will be doing multiple analyses this column name will be repeated again and again . then again a different name. Okay. So, we'll change this to a simpler word. So, that we can easily understand what exactly it is. So, to do that we will be using the colon. So, our colnames are this and we will give this. So, there is already a frequency and we would like to use this cumulative frequency as another column name. So, if I do like this. So, now if I see the value. Okay. So, this is my analysis of frequency and then the cumulative frequency. So, the two analyses are done and similar to the way we have earlier saved the file we can save this file also using the write.CSV function . So, as a practice assignment I request you all to save this file with your name and share with us in the discussion forum. Okay. now if you have noticed here that I have used this camel. kind of writing for giving the column name if you have seen this kind of example that whenever a URL has any space in the names automatically a percentage sign is generated. Okay. So, to avoid this kind of confusion we generally don't use "space extra space" kind of thing in the names. It's always a good practice that you don't give space in the column. Okay. now. So, there is another way of changing the column name because if you have noticed here that you will say that the two columns are already there. Okay. So, why am I typing it again and again while I am writing this? So, is there any option to change only that particular column name. So, this is where we can change the column name of that particular column. So, what we are doing here is that it's a complex kind of function. When we will be discussing the function you will get more understanding about this. So, I request you to try this function after the second lecture when I will be discussing the function. but I will just give you the brief of what I am doing here. So, here I am calling the function colnames and saying that this is the column and where this value is column is available. Okay. and if you have remembered that particular sign. What will it generate? it will generate false, false, false. Why is it generated false, false, false. because there is no column available. with this comes some dots that FQ. Okay. Because I have already changed here. So, first I am checking whether there is any occurrence of column names and then. if any occurrence is there then change that value to a new column name. So, here in this case all these things are false. and after that wherever is true it will change the column name. Okay. So, this is another way I will say it's like an advanced way of changing the column name. and it requires the understanding of both how the function works and then the logical value type we have already discussed. So, I will request that you try this after the lecture. So, now we will move to another analysis where we will calculate the relative frequency. So, how do we calculate relative frequency manually? So, what we used to do first was calculate the sum of frequency. So, the sum is 8 and then we will calculate the relative frequency of 2019 is 1 by the sum is 8. So, it will be 0.125. Okay. and similarly for this relative frequency it will be 0.125. and for this also 0.125 and for this 5/8 is equal to 0.625. Okay. So, this is how

we calculate the relative frequency now we will see how we can calculate the relative frequency in R. So, to calculate the relative frequency in R what we will be doing. we will call this operation and after this operation all the values will be retrieved based on this operation. So, if I do like this if I run this. So, if you notice here that is our relative frequency. So, I can assign like this. So, now if I check the value of relative frequency. So, this is my relative frequency. Okay. So, until now we have calculated the simple frequency, then we have calculated the cumulative frequency, and then we have calculated the relative frequency. So, this is our one analysis and we want to have one single data frame for this analysis. So, what we will do. we'll this is our earlier data frame which has the frequency, and the cumulative frequency. Okay. and I will assign this relative frequency. also with this. Okay. So, now I will assign this. So, this is my relative frequency. So, my final data is this. Okay. So, this is my ear, this is frequency, this is cumulative frequency and this is relative frequency. Okay. So, after that we will export this result as an external file. So, how we will do that we will use the function right.CSV and that is our data frame final data frame. now I will give the file in the location where I want. and I am giving the file users. So, I am just typing the first character and then I am putting the tab. So, it's then where I want to go. if I run this. If I see here there is no file. Okay. The final file is saved in the downloads directory; there is no document directory because I mentioned this part. Okay. So, if I have typed like this. now, if I do like this and if I go here. So, this is my file that was created. Okay. and my earlier file you can see in downloads. So, in my download this is the file. Okay. So, this is how after computing the whole analysis you can export the results to a different file. Okay, until now whatever we have discussed has gone a long way to understanding how things work. but when we have the understanding of what the particular concept is. and how that functions work in R we can easily do the analysis in just a fraction of seconds. Okay. and I will show you that way also. Okay. So, what we have to do is first read the data. So, I am reading again and now what I am doing. I am just tabling the year and then the cumulative sum. Okay. So, here we have to do some modification also like we have to delete some of these columns but in a shorter way in like just two commands we go to the analysis of frequency and the cumulative frequency analysis. Okay. So, this is how you can do this frequency and then cumulative frequency and this then relative frequency analysis in R. I have saved the one practice dataset in our GitHub repository. You give it a try on that dataset for this analysis. and let us know in our discussion forum if you face any of the issues. and if you are getting an incorrect value if you are doing it manually or doing it in R. Okay. So, I will see you in the next lecture where I will discuss how we can create the function and then measure the central tendency. Thank you.