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

Dr. Mohit Garg

Central Library

IIT Delhi

Week: 06

Lecture 24 : Analysis of Bradford and Lotka Laws

Dear Learners, Welcome again. So, in the last 3 lectures we have discussed descriptive analytics. We have seen step by step how we used to calculate descriptive analytics manually then we have seen the descriptive analytics in R also. So, in this particular lecture I will be discussing the two laws which we have discussed in the previous week. One is Bradford's Law and another is Lotka's Law. So, we will be doing the practicals of Bradford's Law and Lotka's Law in R.

So, we have already discussed the theoretical part in the last week. So, as you know, Bradford's Law is one of the popular Law and theoretical distributions for assessing scientific productivity. So, what exactly is it? So, Bradford's Law shows the relationship between the journals and the number of articles published in a subject. So, it was studied by Samuel C.Bradford in the year 1934 when he studied the publication of geophysics and lubrication and he identified some patterns in those publications. What pattern was it? So, the pattern is that if we arrange the journals in decreasing order of their productivity then we can identify three zones of the journals. So, what all these three zones are? So, first is the core zone which we call a nucleus zone where we have a smaller number of journals which have almost one third of the publications. Then we have the second zone which has more number of journals compared to the core journals but it has close or equal to the number of publications but the first zone has. And in a similar way we have a third zone which has more number of journals compared to the nucleus and then our first zone and it also has close to or equal to that number of articles.

We have already discussed these theoretical parts and now we will see them as a practical part. So, for doing Bradford's Law we need a package because we have seen there are a lot of calculations we have to do. So, there are like no inbuilt functions that are there to do this kind of computation but we have some packages. So, that package is bibliometrics so if you do not have the package you can install it. So, we will install the package either by calling the command install.packages and under those brackets we will write the name of the package or by going to the GUI tools and we can install the package also we can search about this package on this URL. So, on the CRAN page we will go here then we will go to this package by name we will search here bibliometrics and this is our package and we can install it. So, after running this thing our package will be installed and after installation of the package we have to call that and to call that package we will be using the library. I will call this package. I hope the package is installed now. We need data to do the analysis.

So, one is that we have to extract the data from the bibliographic data sources like Scopus, Web of science or PubMed. but for the testing purpose we have some dataset. I have already discussed that each package comes with various things like documentations and data also. So, the bibliometrics package also has some data. So, this is that function which we will be using to call the data. So, if I run this, these are all datasets available in the current bibliometrics package. If you want to know about this data function so what we will do we will use the question mark. So, we will use this and help is there so this is the way you can call the data and it is showing what you have to exactly call and what are these documentation.

Then we want to know what all datasets available are with all the packages. So, we call the data function like this where we will say that whichever package has the data list all those things. So, if I run this it is showing that there is a dataset in package bibliometrics. So, these are the datasets, then there is another package bibliometrics data which has these many dataset management, scientometrics and scopus collections. Then we have like in my system there is another package bit 64 then we have this cluster is one package is there like that whatever the package we install the dataset of that package we can see.

So, this was what we have installed in the previous lectures so these are the datasets available and we can use these datasets for our understanding of the analysis. So, if we have to see the dataset of a particular package how we will see it, we will name the package. So, these are the datasets available in this package bibliometrics. Now, for our analysis we need a dataset so there is a package called bibliometrics data and in that package there is some dataset. So, these are some of the datasets we will be using management and scientometrics.

So, we will be using management for the Bradford law and then we will be using a scientometrics dataset for doing the Lotka's Law. So, it is showing here some description also that management uses bibliometric approaches in business and management discipline and then we have scientometrics dataset, co-citation analysis, coupling analysis manuscript. So, make sure that whenever you are running these commands whenever you are calling in this way the package should be called and also the package installed without installation calling of that package has no meaning. So, we can also see these things in our system also where that data is. So, each package comes with a dataset and those datasets are on the folder where the package is installed.

So, we will go to that directory so this is the directory if I copy this and open it here. So, you just copy and paste it here. So, this is the library of all the packages we have installed. So, for us there is bibliometrics and this is a data directory and these three datasets are there then if I go to my bibliometrics data here. So, here we have this data, this management data or scientometrics are there.

So, if you see here that you have just run one single command to install packages and that single command not only installed all those codes but whatever the things like packages and other like dataset and also the manuals also are installed and stored in your machine. So, for Bradford law we will be taking the management data. So, how we will call that is our dataset. This was our dataset in bibliometrics data and under bibliometrics data we will be using management. So, we have to use this data then we have to mention management. So, we are using management and where it is so it is in the package bibliometrics data.

So, I run this so my data is imported here. Now what we used to do first thing after importing the data we will view it. So, we will first view what that data is. So, I will run this here and this is my dataset. If you see this au stands for authors then all those metadata are there then this if we go here this is the DOI then document type whether it's article or see all our article. Some articles are from the conference proceedings.

Then for Bradford law we are interested in the source. So, where our source is yes so this stands for source. Source means our journals so these are the journals if you see here. So, these are all journals for this dataset. We will be using this and if you see here it is showing that there are 898 entries in this dataset and 67 total columns are there. So, with this 67 columns name we can identify the col names and the number of rows also.

So, as a practice you just let us know in the discussion forum how you will be able to identify what all different column names are there and what all different rows are there what exactly the dimension of this particular dataset in R. So, this is our dataset so management is already there. So, after viewing the dataset what we have to do then we will do the analysis. So, to do the analysis we must know what function we have to use, what function names the way we have a search for the maximum minimum and all. So, each package has their own documentation and there they declare that you have to call this function name for doing the analysis.

So, for that if you go on this page this is the reference manual and if you write here Bradford it is on the 12 page. So, we have to call this function Bradford and if you notice here there is only one argument it does not require 2 or 3 arguments we have already discussed the one argument function 2 argument function. So, here only one argument is a bibliographic data frame. So, here our data frame is this management and that is the function Bradford. If I run this our result is here. but I assign to one object. So, I am assigning the whole computation and analysis into an object named Brad result. So, if I run like this Brad result so if I see the output of this graph is there. Now, let us see what exactly the class of this object is. So, this is a list we have already discussed that many of the functions give the output in the list. So, this is a list now to see what exactly the number of elements are there in the list and what exactly their names are. So, what function we use to see the names of the element in the list we use the names function and we will call it like this.

So, this is the name function and here we have to give this object. So, if I have a table and graph so see we have discussed when we were discussing about list that list is a flexible data structure where you can not only store data frame on a vector but here we have stored table and a graph also. Now, for the table we will see what exactly the names are there. So, if I do like this, under the table there are these five columns. One is the SO rank frequency cumulative frequency zone and so we will see what exactly SO is. So, the same way like what we have discussed in the previous lecture, we will be using the dollar sign.

So, if you see this is our list then we will put the dollar so there are two options. So, under the table we want to see that there are multiple options. So, I am seeing SO. If I run it here again all those sources are there so these are 281 then if I have to see the rank so these are the rank and then I have to see the frequency then I can say that this cumulative frequency then the zones. So, the whole dataset is divided into three zones.

If you see only a few of the journals are in Zone 1 then some more journals in Zone 2 and then in Zone 3. So, Zone 1 is our nucleus we will see through our table also. So, now we will convert this as a data frame so we can easily create the table into a data frame. How we will do that we will use the data-frame and this is our table. So, if I run like this, this is the name of the journal and then it is showing what exactly that zone is then let us first create this data frame.

So, our data frame is created now if I run this. So, this is our dataset now there are only five columns. So, one is the source name then the rank then the frequency then this cumulative frequency then what zone is. Now our analysis is done. Now we will export it as an external CSV file. So, how we will do that is write.CSV or read.CSV. No we will be using write.CSV because read.CSV we use for importing the dataset but for exporting the dataset we will be using write.CSV. So, write.CSV. So, this is this thing then we will mention the file name and again here you have to remember about the location of the file. So, my current directory is document only so I am only mentioning that braided result but if you want to save this dataset in another location or desktop or download so you have to give the either the complete path or you have to change the directory to that particular location.

So, our dataset is there now. I will run it like this and my file is created. Now we will see the file here in documents. So, these are the sources and these are the if you see rank and then these are the zones. If you see that only these many 8 journals are there in Zone 1. It has 305 articles. Then Zone 1 has 8 journals and the number of articles is 305. Now we will see Zone 2. So, in Zone 2 we have 65. So, in Zone 2 we have 65 journals and then we have 604. So, 604 minus 305, 299. Then we will go to Zone 3. So, in Zone 3 how many journals do we have? 208. And the number of articles we have? So, it has a total of R898. So, 898 minus this 2 will be 294. So, I see here this total is 898. So, these are our 3 zones. So, Zone 1 is our nucleus. This is our nucleus. These 8 journals are our nucleus. Then if you see here that Zone 1 has the least number of journals, but one third of those articles. similar way Zone 2 has more number of journals. It is close to 305 only and then we have in Zone 3 208 journals. So, this is what a Bradford law is. This is exactly the Bradford distribution.

So, now the question comes that whenever we discuss R and its packages we say that R is an open source and R packages are open source. What exactly does open source mean? So, the open source means that it is available for free, but your code is also available. You can edit the code. So, here what we have done? This function is already created. If you remember when we have created the addnum function or mohit function or Mukesh function, we have created and we have given the command. We have to print this, but this function is the main thing which calculates the whole analysis. So, we have to see how this function is doing all the computation. And because this package is open source, all that code is available.

How will we see that? So, we will go there. So, this is our code page and here if I go on under R, I can see there are a lot many. What computations are there? Those functions are there. So, what we are doing is we are doing for Bradford. So, if you see this code, they have used the comment option for this hash. So, it is showing that this is a Bradford law and the formulation is that if a 4 journal in a field are sorted by number of article into 3 groups, each with about one third of all articles, then the number of journals in each group will be proportional to $1:n:n^2$.

So, this is our code. This is our function. So, we have created only one or two argument functions and with the two or three lines, but here if you see there are many lines. So, they have developed this wonderful package and they have written the whole code and this whole code is free. You do not have to pay any kind of amount to access this. Only thing that I mentioned in the starting is that whenever you are using R or its package, you give the citation. At least acknowledge that they have done this much contribution and because of them, we are able to do this analysis.

So, we will understand a bit about this code because if you see here, this Bradford function just generated this graph and in one way you can just simply export this graph save as image or this R code 04 and say for example, I name it as Bradford. So, it is created, but you see it is black and you want to change the color and or maybe this heading or anything. So, that function does not provide anything. Means you cannot assign that, you change the color or you can assign your different font size. So, for that you must understand how that function is working.

I am not saying that you must go through the whole code, but some of the key things which will be helpful in designing this visualization in a better way. We will be learning about data visualization in the upcoming weeks, but just see how exactly if we understand this function code, we can easily change this whole visualization. At least we can change the color or the legend and all. So, I am copying this code. So, I am copying this code and after copying it, I am pasting here.

Okay, so I have copied that whole code and I have named it as a new function name. So, I am just making it look like a new Bradford look. And now this is my function. Now on the same data set, we won't call Bradford, but we will call it a new function. What name have we been given? New Bradford. So, if I run this, we have not created the new Bradford. So, it automatically pops out when first you create the function. So, what exactly is our new Bradford? So, new Bradford is this. Now if I run this, I will assign it to an object. Could not find function AES. So, it's because we haven't called another library, library ggplot2. We will call the library ggplot2. Now if I run this and now I hope and now if I run this, I hope it won't show any error. Okay, so the thing is created and this is our new Bradford results. So, this new Bradford result is a new object which has the output of a Bradford computation, but based on the new function. It's the same function. We have just copied, but only changed the name of the function, but the results are the same.

But now we will make some changes in this function so that we can change some color and all those things in this graph. So, if you see the class of these new Bradford results, it's a list with the same output as what we got there. Now if I like to change the color. So, here if you see black is written. Instead of black if I write green. Now I'll run it again. Now if I run this. Okay, if you are able to see this the color is changed. So, this is a green color, earlier it was black if you see this was a black color. So, this is another important feature in RStudio that you can just go around and see what was the previous visualization and what is the current visualization. Okay, so this is our current visualization. So, we have changed this to green now, maybe now you want to remove this logo. Okay, so but I recommend that whenever you are using this code and this package, please don't remove the logo. At least it's a small contribution to their work. What wonderful things they have done to develop this package.

So, if I do like this if I do like this and now if I run this again. Okay, so this logo is gone you can put your own logo or something else is there, but do cite this particular package if you are using this whole code. Now if you want to change another color of maybe so these are what exactly mentioned here is the hexadecimal code. Okay, instead of the color name like green red blue you want to have some shades of that color. Okay, so you can use those hexadecimal codes and for that you can go to this page HTML color codes and here so if you see our color was this triple 4. So, first we will see what was so our color code is 6 times 4 not triple 4.

So, I will just mention here that this is our color now. I think we will change it to some other color, maybe blue shade and we will make a difference. Okay, so we will copy this here and we will paste it here. Okay, so if you see here that this is again like the beauty of the R studio interface that whenever you are putting the hexadecimal code it automatically identifies the color and you also get the idea okay you are using this color. Okay, now if I run it again here is this code.

Okay, so that color was for the core sources for Bradford's Law. Now if I make some more changes in this, let's say for example we want to add something like here we want to add Bradford's Law NPTEL. Okay, let NPTEL Bradford's Law. Okay, then what we want to change is the more colors. So, this white and let's see if there are any other colors there. This is white or maybe instead of this black we will use it as maybe cyan color and now if I run this.

Okay, so if you see here what all things we have changed we have changed this text then we have changed this y-axis. We have also changed the x-axis color. So, if you see our earlier plot, it has this text and then this x-axis was in black color before we have this logo. We have also removed the logo and this black color was there. Okay. So, this is how you can play with this whole code and that is why this whole package is open source and it's totally free.

Okay, so I recommend that whenever you are using Bradford's Law or any of the computations in R or using its package don't go with these three commands. Okay, what all those things we have done here. So, if you see here the whole Bradford's Law computation can be done with this single function. Okay, and you can just simply use it for your analysis, but if you want to make it more beautification and if you want to add some more things and some more extra information into that. Okay, so for that you have to use this code and you have to do some little changes in that.

Okay. So, this is all about the Bradford Law computation. Now, we will discuss the second law which is Lotka's Law. So, Lotka's Law is another popular law of scientific productivity. It was given by Alfred J. Lotka who studied the author's productivity. So, what exactly he studied is that the number of authors making n publications is about 1 by n square of those making 1 and the proportion of all the authors who are making the single contribution is about 60%.

Okay, so we will see how we can compute in the R that Lotka's Law. If you are doing this Lotka's Laws computation in continuation of this Bradford's Law, you don't need to call this bibliometrics package, but if you are doing it from scratch and your session is new, this package will call it. So, R package called. Now, we will go with the data. We have to see the data. So, to see the data we will use this data function and if I run this, these are all datasets in this bibliometrics package.

Okay, and then there are like different other datasets in different packages. Now, we have already seen that if you are a beginner, you must try each time whenever you are using any of the laws and you are using any of the test datasets. Okay, so you can do this now if we go here. So, in bibliometrics data there are like different datasets where we will be using this scientometrics. Okay, so this scientometrics is a dataset of a journal scientometrics. So, we will be using this particular dataset and now if I run it here. Again, the same thing our data is imported. What will be the first thing we will do? We will view the dataset. So, if we click here. So, this is our dataset. This dataset has 17 columns and then it has 147 entries. So, then if you see here, it is the source of all scientometrics because it is a dataset of a scientometric journal and document type is also article and mostly like some of them are proceeding papers, but articles are there. Then it has those titles and other things are there. Now, after we have viewed the data, we have got to know what exactly this data is and where we have to calculate this. So, till now in the Bradford law there was only one function Bradford which calculated your all Bradford law analysis.

Okay, but here in Lotka's Law in this particular package it has two functions. Okay, so we will see the first function and then we will see the second function. So, that first function is the biblioanalysis. Okay, then if I run this here and this is our package.

So, we will put it like this. So, the first part of our Lotka's Law is done. So, I am assigning this to an object. Now, our second function is this Lotka. Okay, so I am calculating like this. So, this our whole Lotka computation is in the object Lotka result. So, we will see the first class of this particular object. If we see here, it is a list again the same way the output of Bradford is. Now, we will see what all names are there in this list element. So, here it is like there are six elements. Here there is no graph. Okay, so now we will go one by one if I see the results. This is the result you see here of the number of authors with a single contribution or more compared to these two contributions and three contributions. So, this is what the Lotka's Law is and now we will see each element one by one. So, I will just run this. So, again here we are using the dollar sign to access the one particular element of that list. Okay, so how we are using it Lotka results and then dollar and then author productivity and if I run this here.

Okay, so this is the output then this is the beta value then this is C then if we see this. So, this is our this value then fitted value then P value. Okay, so our whole computation of Lotka is easily done with these two functions. What we have done is we have used this Biblio analysis and then after that we have made the Lotka's Law on this particular output. So, here let us now again see the functions of Lotka’s.

Okay, if I go here to Lotka this is a Lotka. Here the old details are given what this Lotka function calculates. So, the Lotka's Law estimates the Lotka's Law coefficient for scientific productivity. So, we have already discussed theoretically in the previous week how exactly the Lotka law is and now this function is okay. So, this is the Lotka function but before that we have used the function Biblio analysis.

So, if you do Lotka's Law on our that particular data set it won't work on Scientometric. Okay, or maybe another dataset. So, first you have to use Biblio analysis and then you have to use the function of this Lotka. So, before this we will go to the Biblio analysis. So, our Biblio analysis function is here. So, it is shown that in the Biblio analysis this particular function performs a Bibliometric analysis of a dataset imported from Scopus and Clarivate analytic Web of Science databases and then M is a data frame.

So, for the Biblio analysis function we require only one argument. Okay, that is M that is a data frame and all those things are there that authors and all those what all data it require and this is a big function. Okay, so the same way we can copy this code and we will paste it here and so this is our first function, a Biblio analysis we have copied the whole function. So, M is the Bibliographic data frame and these are references. So, we are not changing anything into the first function because that is a basic function of this particular package Biblio metric and in Brad Ford because we have to directly do the analysis on the dataset.

So, we were able to do some changes and we can change the visualization and all those things. Okay. Now we will run this and this time I am naming it as results if I run this. Okay. Let us name it as one. Let us see how if we copy paste will it work. The same mistake, so this is a common mistake whenever we are copying the code and when we are running it and sometimes that without creating the function we are calling that function.

Okay. So you have to remember that whenever you are calling the function you must first create it. Okay, that is why it is showing that this error is coming. So we haven't run that function. So first we will create the function.

Okay. Now if I run this again see so this is a function in a package called as the package. So we will call that package also. And now if I run this again. It is done.

Now if I like this was our another function for Lotka. So what was that? So this was a Lotka function. So I am just copying it and pasting it here and naming it as a new Lotka. I am running it now and then if I see the output of this new Lotka function and if I see the same output. Okay. So this is how you can use the code of the package whatever the function they have created and you can copy it and edit it according to your need but you should remember that whenever you are doing like this either you are editing or you are using the package you must give the citation to that particular package or that particular software.

Okay. So if I see here on this page also this license and details are given. So if you see the license here, this particular program is like free software. You can redistribute it or modify it under the terms of GNU General Public License as published by the Free Software Foundation. Okay, so this is all about two of the laws that are Bradford's law and Lotka's Law, another law which is Zipf's law. I will be discussing when I will take up text mining and because the Zipf's law is related to the frequency of a text at that time. I will discuss that particular law recommend that you use this Bradford law and change the color or maybe something text and please let us know in the discussion forum if you like face any kind of difficulties in doing that we will be very happy to resolve those things or if any new thing you come to know, please do let us know. Okay. So thank you. We'll see you next week. Thank you.