Probability Foundations for Electrical Engineers

Prof. Andrew Thangaraj
Department of Electrical Engineering
Indian Institute of Technology, Madras

Lecture – 26
Real-life Modeling Example

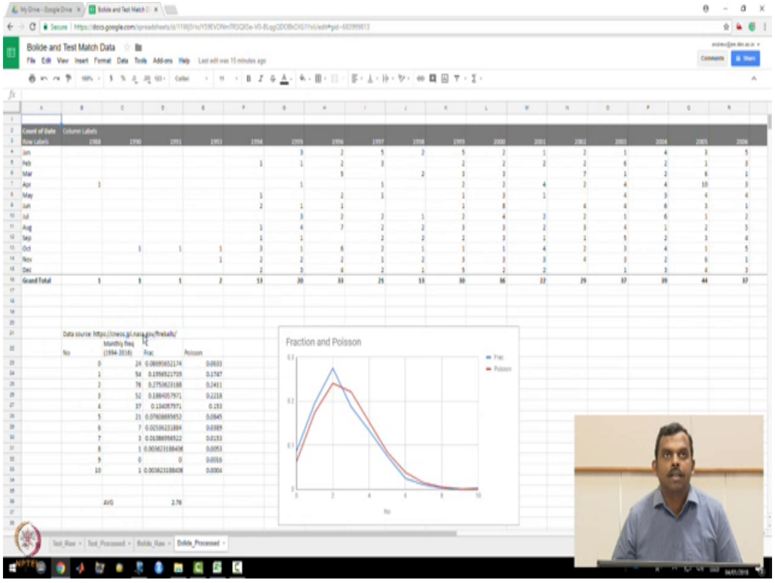
Welcome to this lecture on an interesting modeling example. I am going to show you 2 sets of data and in some ways of modeling them in a probabilistic way. So, now, we have been studying about the probability random variables independence events and all of that distributions PMFs all these are valuable tools when you want to model a real world phenomenon using probability.

So, in this short video I will show you 2 different data sets data that I have taken from the internet I will also give you the source and with this data I am going to try and fit the probabilistic model and you will see in one example the model will be a very good fit, in the other example the model will not be a good fit and there is a good reason for that and I will show you why that is. So, and I am going to be using the Poisson model and if you remember from professor Aravind's lectures the Poisson model applies when you have a interval of time and events happen in that interval of time at a specific points and you are interested in the number of events in a particular time window.

So, the examples I am taking are the first one is going to be meteor strikes on earth. So, many of you might know this phenomenon earth is in space and earth's atmosphere protects us from meteorites that constantly get pulled into the earth by the gravitational attraction, forces a lots of small bodies of different sizes and shapes which enter the atmosphere and this get burnt out in the atmosphere. So, a significant meter meteorite or big enough one is called a bolide and this bolide data you can get from NASA websites.

It is freely available maybe I should show you on the screen the actual URL.

(Refer Slide Time: 02:07)



Bolide and Test Match Data

Count of Date	Column Labels																
Row Labels	1988	1990	1991	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Jan						3	2	5	2	5	2	1	2	1	4	3	5
Feb					1	1	2	3		2	2	2	2	6	2	1	3
Mar							5		2	3	3		7	1	2	8	1
Apr	1					1		1		2	2	4	2	4	4	10	3
May					1		2	1		1	3	1		4	3	4	4
Jun					2	1	1			1	8		4	4	6	3	1
Jul						3	2	2	1	2	4	2	2	1	6	1	2
Aug					1	4	7	2	2	3	3	2	3	4	1	2	5
Sep					1	1		2	2	2	3	1	1	5	2	3	4
Oct		1	1	1	3	1	6	2	1	1	1	4	2	3	4	1	5
Nov				1	2	2	2	1	2	3	3	3	4	3	2	6	1
Dec					2	3	4	2	1	5	2	2		1	3	4	2
Grand Total	1	1	1	2	13	20	33	21	13	30	36	22	29	37	39	44	37

Data source: https://cneos.jpl.nasa.gov/fireballs/

No	Monthly freq (1994-2016)	Frac	Poisson
0	24	0.08695652174	0.0633
1	54	0.1956521739	0.1747
2	76	0.2753623188	0.2411
3	52	0.1884057971	0.2218
4	37	0.134057971	0.153
5	21	0.07608695652	0.0845
6	7	0.02536231884	0.0389
7	3	0.01086956522	0.0153
8	1	0.003623188406	0.0053
9	0	0	0.0016
10	1	0.003623188406	0.0004
	AVG	2.76	

You can see that right there that is the data source it is a jpl nasa dot gov site slash fireballs you can go visit that and get this data and the data will be in a raw format first let me show you what the format of the data will be it will have the date and time of the meteorite striking. The latitude and longitude where it struck and the altitude at which it may be burned down the velocity at which it came out and few other data you can go to the website to see what it is and some impact and energy etcetera.

So, I am going to be interested in. So, this is our event. So, we have an interval of time let us say from certain year to certain year and within that year every month there would have been a certain number of meteorite strikes. So, now, I am interested in modeling that number of meteorite strikes per month as a Poisson distributed random variable. So, now, you can think about when the Poisson approximation is a good approximation or Poisson model is a good fit you might remember from Professor Aravind's lectures that what happens when one interval of time and what happens in another interval of time which does not overlap with the first interval has to be independent. If it is independent then the Poisson model is a good fit and if you think about bolide or bolide or meteorite strikes the what happens in one month and what happens in the next month is going to be independent it is not really closely connected. So, we expect a Poisson model to be a good fit.

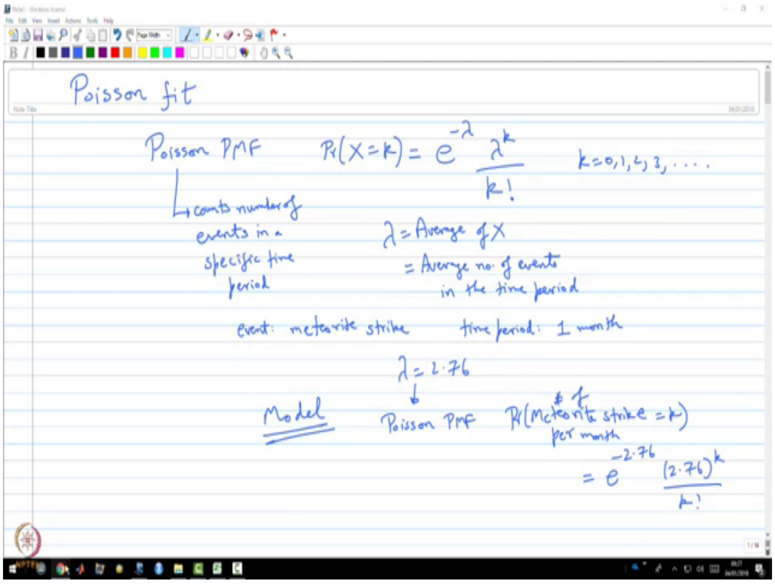
So, I have taken this raw data and I processed it a little bit and the processing that data basically was I organized it year wise 1988 all the way till if I am not wrong 2016 and every year I divided it up into months you have January to December. So, I have started, I counted the number of strikes which happened let us say in June 1994 that is 2 in number. So from the data from the NASA website you have to organize it in this format. So, this requires some minor scripting and a playing around with excel, excel itself is good enough depending on your comfort you can do this.

So, you see various numbers here for instance this month in August 1996 has seen 7 meteorite strikes etcetera etcetera. So, this is the data month wise for meteorite strikes. Now, we have to take this data and then again further process it and assemble it in this fashion. So, what I have here in this table that you can see here there are 4 columns here first column I have called as number this is the number of meteorite strikes per month, so 0 1 2 all the way up to 10. So, the second column here I called with this monthly frequency. So, this is the way to understand this there were 0 meteorite strikes for 24 months. So, remember how many months total number of months I am considering 1994 to 2016, that is 23 years, 23 into 12 is 276 months.

So, I am looking at a total of 276 months. In these 276 months, 24 of the months has 0 meteorite strikes 54 of the months had 1 meteorite strike, 76 of the months had 2 meteor strikes and so on. So, that is the way to understand this and this fraction computes the fraction of months in which there were 0 meteorite strikes. So, this is 24 divided by 276. This is a fraction of months in which we had one meteorite strike, so that is 54 divided by 276 that is the fraction. So, you keep on getting this fraction. And then this is the average number of meteorite strikes per month. So, this you have to compute as 24 into 0 plus 54 into 1 plus 76 into 2 plus so on divided by 276 that is the average number of meteorite strikes per month that that ends up being 2.76 over. So, this is the way in which you compute this data.

This is, at this point this is only data I have only tabulated it just looked at the data and counted it in different ways figured out, how many times, how many times, how many months we had 0 strikes how many months you had one straight etcetera. So, now, we have to do a Poisson fit to this model let me go to my writing page here. So, I am talk about Poisson fit.

(Refer Slide Time: 06:31)



Poisson fit

Poisson PMF $Pr(X=k) = e^{-\lambda} \frac{\lambda^k}{k!}$ $k=0,1,2,3,\dots$

↳ counts number of events in a specific time period

λ = Average of X
= Average no. of events in the time period

event: meteorite strike time period: 1 month

$\lambda = 2.76$
↓
Model Poisson PMF $Pr(\#\text{ of Meteorite strike per month} = k)$
$= e^{-2.76} \frac{(2.76)^k}{k!}$

Now, what is the Poisson distribution? If you remember the PMF for Poisson distribution probability that X equals k is e power minus lambda lambda power k by k factorial what is lambda, lambda is the average of x so in some sense the average number that you are interested in. So, now, remember professor Aravind's lectures he was mentioning how the Poisson PMF. So, k equals 0 1 2 3, so on. He was mentioning how the Poisson distribution actually counts the number of events can be modeled to count the number of events in a specific time period right and this lambda will then become the average number of events in that time period.

So, what is it that we are interested in? Now, the event that I am interested in is meteorite strike and the time period I am interested in is one month and I have gone ahead and collected the data for k equals 0. So, that is the tabulation that you saw here I am sorry that is the tabulation that you see here. So, 0 corresponds to k equals 0 and the fraction of times that that happened which is roughly the probability of k equals 0 is this 0.0869 so on all the way down to let us say for instance 6 k equals 6 this number of meteorite strikes in a month is 6 and the probability of that happening is 0.0253.

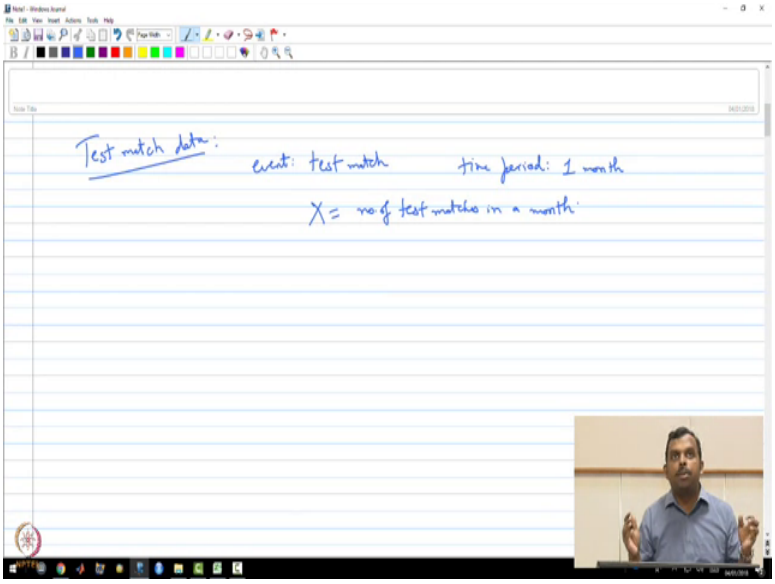
Now, I am roughly interpreting it as the probability, I am roughly interpreting this as a probability, but what is it actually it is the actual fraction of times that event happened in that time period. So, remember again my time period is 1 month within that month I am looking at how many in what fraction of months I had 6 meteorite strikes and that gave me this and I also have the average. Now based on this lambda, my average values lambda at lambda is 2.76 based on this you can get a Poisson distribution. This gives you a Poisson PMF and here probability that meteorite strike equals number of meteorite strike per month in a month right this k will actually be e power minus 2.76, 2.76 power k divided by k factorial.

So, this is my model. So, this is my model remember, this is my model this is my probabilistic model the actual data came from the NASA website and then I fit a probabilistic model and these are the numbers that are calculated there this is what I am calling as Poisson fit. So, for k equals 0 you get 0.0633 for k equals 1 you get 0.1747 and you can see the close relationship between this and this you see 0.19, 0.17 fairly close next is. So, k equals 2 you see it is fairly close fairly close and so on.

You can actually make a sketch of these two things the fraction and the Poisson fit versus the number from 0 to 10 then we see it is a reasonably good fit. So, we have the same sort of behavior. So, this is how people work with probabilistic models in practice you go look at the data and then also have a good understanding of what kind of distribution will model the process by which this data was collected or the phenomenon involved in this process that know that is very important. If you model that wrongly you will get a very bad fit it will not be a good fit and you also saw that the fit is only approximate you will never have an exact fit you know. So, that is always true in fact, all models all theories work like that even deterministic theories now will give you an exact fit you know in practice its always an approximation. So, probability also works like that you will get an approximate fit most of the time and you can see how this was work. So, this is the bolide data event the event being bolide hits on the earth meteorite strikes on the earth.

The next data I am going to consider is from test matches. So, here the event I am interested in is a test match and we will keep the time period as once again its one month.

(Refer Slide Time: 11:21)



Test match data:

event: test match

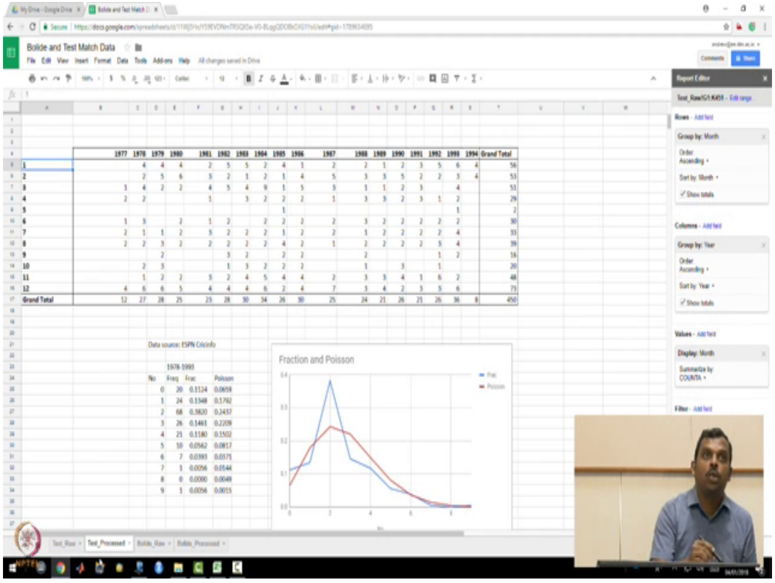
time period: 1 month

X = no of test matches in a month

So, I am interested in how many matches test matches I will of course, this match refers to cricket test matches how often test matches how many times test matches are held in one month. So, that is my X that I am interested in, number of test matches you know.

So here again you first need to collect data over a period of time count the number of test matches that happened in months, I have done that as well. So, let me show you where I got that from.

(Refer Slide Time: 12:11)



	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	Grand Total
1		4	4	4	2	5	5	2	4	1	2	2	1	2	3	5	6	4	56
2		2	5	6	3	2	1	2	1	4	5	3	3	5	2	2	3	4	53
3	1	4	2	2	4	5	4	9	1	5	5	1	1	2	3		4		53
4	2	2			1		3	2	2	2	1	3	3	2	3	1	2		29
5									1								1		2
6	1	3		2	1	2		2	2	2	2	3	2	2	2	2	2		30
7	2	1	1	2	3	2	2	2	1	2	2	1	2	2	2	2	4		33
8	2	2	3	2	2	2	2	2	4	2	1	2	2	2	2	3	4		39
9			2			3	2		2	2		2				1	2		16
10		2	3			1	3	2	2	2		1		3		1			20
11		1	2	2	3	2	4	5	4	4	2	3	3	4	1	6	2		48
12	4	6	6	5	4	4	4	6	2	4	7	3	4	2	3	3	6		73
Grand Total	12	27	28	25	23	28	30	34	26	30	25	24	21	26	21	26	36	8	450

Data source: ESPN CricInfo

1978-1995

No	Freq	Frac	Poisson
0	20	0.1124	0.0659
1	24	0.1348	0.1792
2	68	0.3820	0.2437
3	26	0.1461	0.2209
4	21	0.1180	0.1502
5	10	0.0562	0.0817
6	7	0.0393	0.0371
7	1	0.0056	0.0144
8	0	0.0000	0.0049
9	1	0.0056	0.0015

So, first I should believe acknowledgement of where I got the data from the data I collected is from espncricinfo you can you can also go and go to that page and collect the data. The datas in this form it give you the runs, wickets, balls, average run per over, you can calculate, you can analyze any other data that you like and do fits of different models that is up to you, but the more data I am interested in is how many test matches were held in a particular month that is what in every month.

So, this gives you the list of all the matches where it was held start date in the year etcetera etcetera. So, from here I have processed this data a little bit and from 1977 to 1994, I have now tabulated it in a different way. This is the months 1 to 12 is the month January, February, March, April all the way till December. This is the year and for instance this number 3 means in July 1981 there were 3 test matches and this number one means in August 1987 there was one test match and blank means there were no test matches now. So, this is the data you can look at this data again and then once again you have to now compute and convert this data into a form that we can use. So, that is the form. So, this is the frequency count. So, similar to much like what I did before. So, it means what is this frequency count means for 20 of the months that I considered between 1978 and 1993, once again you can see how many months they have considered from 78 to 93 that is I believe 16 years. So, that 16 into 12 would be 192 right. So, that is the number of months that you would have here.

And then you can divide 20 by 192 and you will get this 0.1124. So, that is the way in which you figure out this fraction. So, (Refer Time: 14:02) is the fraction of months in which there were 0 test matches. So, likewise you have one test match, there were 24 months in which that happened that is the fraction 2 3 4 5 6 7 8 till up to 9 and you have these fractions and the average number of test matches its roughly about 2.72. So, that is the interesting number. So, sometimes you should not be looking at wrong statistics to compare. So, a number of average number of meteorite strikes on earth is 2.76 per month and average number of test matches in the that happen is 2.72. So, can we conclude that test matches and meteorite strikes are related not really you know; I mean do not do not go into the wrong conclusions, but anyways it is interesting to see that the average is roughly the same.

So, once again I did a Poisson fit exactly like before I just took this average and did a Poisson fit on this and if you make a plot here you see that the plot is pretty bad you know it is not very close to the fraction. So, on the meteorite the data we got a wonderful Poisson fit for the event which is a meteorite strike per month. Now, if I do the different event which is test matches per month I get a very bad fit, the Poisson fit is a very shallow and in some sense, but the test match fit ends up being very peaky around to it is its peaking a lot a lot of months when they tend to happen they have to test matches.

Now, if you think about it there is a very simple explanation for why the fit is bad here, the reason is if you if you consider test matches how do test matches happen usually one country visits another country right. So, they go on a tour and then they play a series of games maybe 3 games, 5 games, 5 test matches and they all happen one after the other. So, if I told you that India played Australia in March first 10 days of March 2005 or something like that then it is very very likely that India played one more match with Australia in the remaining of the month. So, it is not going to be independent how many events that happen in a particular time period are not going to be independent of a number of events in the next time period because the country is touring another country and they play one test series within a month. So, the Poisson model which assumes independence of events over different intervals of time is not a good fit when you come to test matches.

On the other hand for meteorite strikes which are random events from space that ends up being a good model and you get a good fit. So, if you want to model the number of test matches that happen in a particular month you cannot use a Poisson model you have to use something else. Now, what model you want to use that is up to you if you find a more interesting model that fits this description then you might want to use it, but at least Poisson it is not a good fit.

So, that concludes this brief lecture on fitting data and probability models. Hopefully you learned a couple of lessons on how to take data and how to fit it into a probabilistic model and how to see if the Poisson model is a good fit and how to argue based on your knowledge of how the Poisson's model is derived to see which case it applies well in which case it does not apply.

Thank you very much.