Data Analysis and Decision Making – I
Prof. Raghu Nandan Sengupta
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 28
Data Properties

A warm welcome to all my dear friends and students. This is the DADM which is Data Analysis and Decision Making -I course under the NPTEL MOOC series and this course as you know is for 12 weeks, 30 hours, 60 lectures and each week we have 5 lectures each being for half an hour duration. So, as you can see from the slide we are in the twenty eighth lecture; that means, we are into the sixth week. So, 30 lectures would basically complete the sixth week.

Just a few things which I thought it is almost in the middle of the course I thought I will let you know, technically we would have been best that if I had shared all these things in during the thirtieth lecture. But, this is a query from my part that how interesting the course is for you, for all those who are taking this course. Whether you have understanding this and whether you are you are able to appreciate the content and you have checked the books which have been suggested. Because the reason is that that many of the queries which are coming in the forum are based on things which are maximum of they are, are already discussed and the explanations have been given.

And, also many of the queries which are arising are of see for example, for the first week. So, we as a team from NPTEL side myself as a instructor and the TAs we want to basically enquire from your end if how things are there how things how good or bad this course is going on, whether you are understanding it, whether the concepts are clear because what in a 30 hour 30 hour lecture where you have assignments 6 weeks. We will be in the 6 weeks or 6 weeks of assignment would be coming up as you complete the 6 weeks.

So, in the 12 weeks you will have 12 assignments, you will have question papers accordingly. So, I think it is pretty well spread over where you will be able to read the concepts answer the questions and clear your doubts. But, main thing obviously, happens is that in this duration of 30 hours to basically give the whole gamut of statistics from the data analysis and decision making point of view may not be possible. So, I will urge that students please read the books which I have suggested. They are I am not saying that the classic books, but they are adequate enough which will basically clear your doubts.

And, then if you can basically ask questions obviously, you have full right to ask questions on what is being taught plus you can ask questions related to the books giving the references it will be easy as for us to clarify the points. Because maximum the queries which are based to or the fact that they are little bit older discussions which we thought you would be immediately raised or they are queries that which may not be directly pertinent to the course. So, this is an honest request from my part and I am sure you will be able to take care of that and will be able to deliver the course in the coming 6 weeks in much much better way.

So, as you know as I am do mention each and every time; so, we are in the sixth week and I am Raghu Nandan Sengupta from IME department IIT, Kanpur with this teaching this course. Now, the issue what we were discussing in the last class which was in the twenty seventh one was to do with the concepts of forecasting, the concept of trend analysis concept of seasonality, cyclicity and the models of holt linear, holt winter. They are the one of the so called adequate models which are able to take care of for the trend seasonality average movements and all these things. Plus the concepts which I have discussed for regression part, they are also able to take care of your queries in a you know better way.

Now, I did mention that I will be discussing the problem from multiple linear regression, but multiple linear regression would be more from the point of view of multivariate statistics. So, I will come to these concepts of problem solving for multiple linear regression a little bit later. So, this is a slight bit change in the overall plan, but it would not basically stop the flow. The flow continuously remain, it remains the same; there would not be any loss of knowledge neither loss of the break of the continuity or the break of the concept which you are going to pick up. It is just a different change or where it should be adequately placed.

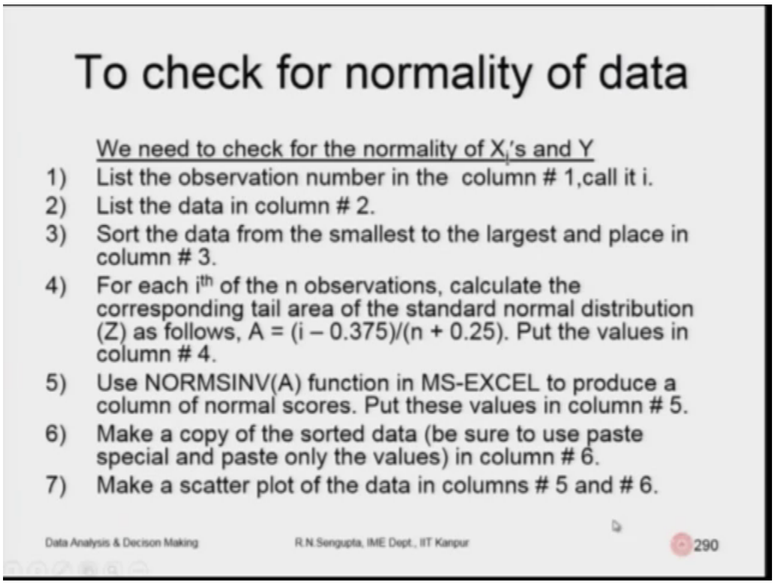
Now, in the last part of the slide which I will again discuss today because generally whatever I slide I do for any class or any lecture I complete that and then go to the next slide, in the next day or in that same lecture, but this slide would be a little bit important because we will be discussing about normality in greater details.

Now, if you remember the concept of normality was stated by me during the course when we are discussing the normal distribution x being normal with mean mu and sigma square variance. Then we also discussed that how you could convert x which is a normal distribution to the z distribution case standard normal and we also considered that later on that considering the central limit theorem to be true. We can consider any distribution with its mean and variance and plug it as the mean and variance. Not plug it literally, they would be conceptual change how you do that if you remember the mean values of that particular distribution variance of that particular distribution that they sample counterparts would be utilized and we will basically be able to solve this accordingly for both the interval estimation problem as well as the hypothesis testing problem.

Now, for the normality test generally we have a very simple concept which is known in the Q-Q plots which is the quantile-quantile plots. The quantile-quantile parts concept is like this. As required I will be writing many things drawing many diagrams as I generally do for different slides when I am discussing. So, the quantile, for the quantile-quantile plots the issue is that you want to basically plot two different distributions, one being the standard normal other being the unknown distribution. Compare the quantile-quantile plots and basically say whether that particular distribution which you are trying to compare with the standard normal distribution is normal or is skewed to the left or skewed to the right.

Now, whether it is skewed to the left or the right would come up very nicely when you have the diagram for the Q-Q plots and when I discussed that accordingly. Now, the Q-Q plots conceptually is that you basically plot both the distribution as I say and you and you first rank them and then plot. So, what is going to happen, I am going to basically discuss it accordingly. So, we need to this is the slide which you have missed which you are not missed we have did covered, but we are going to cover it again in order to basically clear the doubts in much better fashion.

(Refer Slide Time: 08:30)



To check for normality of data

We need to check for the normality of X_i's and Y

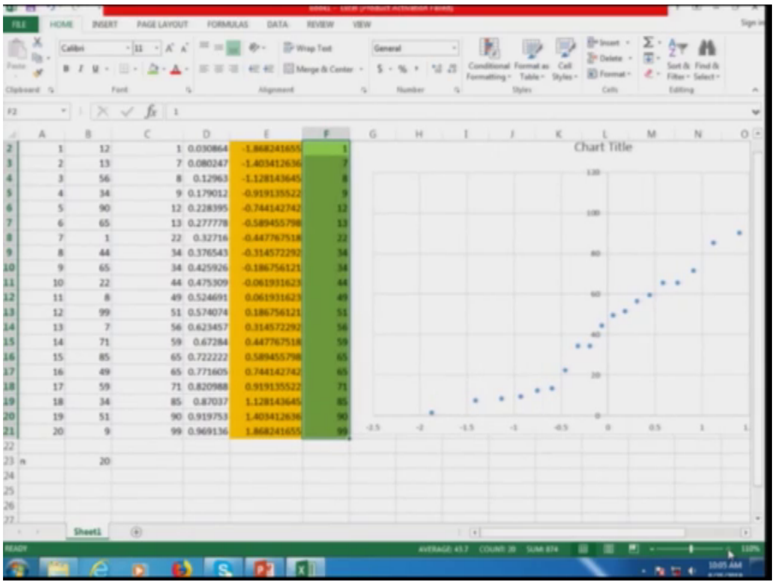
1) List the observation number in the column # 1,call it i.
2) List the data in column # 2.
3) Sort the data from the smallest to the largest and place in column # 3.
4) For each i^{th} of the n observations, calculate the corresponding tail area of the standard normal distribution (Z) as follows, A = (i – 0.375)/(n + 0.25). Put the values in column # 4.
5) Use NORMSINV(A) function in MS-EXCEL to produce a column of normal scores. Put these values in column # 5.
6) Make a copy of the sorted data (be sure to use paste special and paste only the values) in column # 6.
7) Make a scatter plot of the data in columns # 5 and # 6.

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 290

So, we need to check for the normality of Xs and the normality of Y and one of them we consider to be at the standard normal deviate and based on that we will proceed. So, you list the observations in the column number one. So, you list the number 1, number 2, number 3. So, that is basically the serial number. So, let me try to do it accordingly. So, this is the first time I am trying to basically do the problem accordingly. So, it will be easier for all of. So, I will switch over from the ppt slide to the excel sheet for better understanding.

So, I want be going to the view mode, please excuse me and please appreciate how I proceed. So, list the observation number in the column 1 and call it i.

(Refer Slide Time: 09:01)



F2 | 1

	A	B	C	D	E	F
2	1	12	1	0.030864	-1.868241655	1
3	2	13	7	0.080247	-1.403412636	7
4	3	56	8	0.12963	-1.128143645	8
5	4	34	9	0.179012	-0.919135522	9
6	5	90	12	0.228395	-0.744142742	12
7	6	65	13	0.277778	-0.589455798	13
8	7	1	22	0.32716	-0.447767518	22
9	8	44	34	0.376543	-0.314572292	34
10	9	65	34	0.425926	-0.186756121	34
11	10	22	44	0.475309	-0.061931623	44
12	11	8	49	0.524691	0.061931623	49
13	12	99	51	0.574074	0.186756121	51
14	13	7	56	0.623457	0.314572292	56
15	14	71	59	0.67284	0.447767518	59
16	15	85	65	0.722222	0.589455798	65
17	16	49	65	0.771605	0.744142742	65
18	17	59	71	0.820988	0.919135522	71
19	18	34	85	0.87037	1.128143645	85
20	19	51	90	0.919753	1.403412636	90
21	20	9	99	0.969136	1.868241655	99
22						
23	n	20				

So, this is A-th column and consider this is i. So, I have the reading number. So, reading number consider I plot it. So, there is 1 2 3 4 5 6 or this values are given. Now, list the data in column 2. So, we do not have any data. So, we consider let us take if we have let us assume some data (Refer Time: 09:49) data and we will use that data and corresponding to based on that we will try to see. So, consider the data is I am putting an hypothetical arbitrary data; this will become clear as we do it later. I should do it till 20, let me make it very simple for ease of understanding .

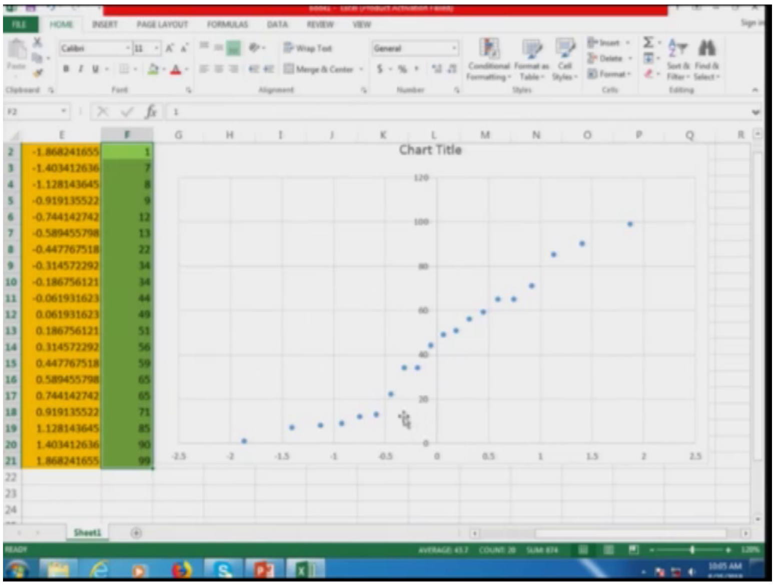
So, this is the data which you have. So, this is let me write it the data. So, this is column 2. Sort the data from the smallest to the highest and proceed in the third column. So, what we do is that copy it here first now you select it. So, you go to sorting, smallest to the highest you continue with the current list and this is basically the values. So, data sorted one (Refer Time: 11:40).

Then you for each i-th of the n of the n observations; so, n you know n is basically 20. So, let me write it down. For each of the i-th observation n observation calculate the corresponding tail area of the standard normal deviation following this formula i minus 0.375. So, this would be is equal to i means it will be the data number which is here minus i minus, ok. So, this would be i is here minus 0.375 divided by n is let me make it n coming from here plus 0.25. So, what I do is that because sorry. So, this is has to be building on dollar sign because that is fixed. So, I copy it all the values. So, just check.

So, the formula remains the same just check where the, I am just highlighting it now to make you understand that we have done it correctly.

So, now, use NORMSNIV A. So, to produce a column of normal scores put these values in. So, A column is basically this is the data you have basically found A. So, let me put it A. Now, put as NORM NORMSNIV A. So, let me increase it. So, this is A is equal to put the values in column 4 done NORMSNIV A. So, NORMSNIV A is this NORMSNIV. So, this is the value which I will take put it. So, I copy it put this value common mark a copy of the sorted data to be sure use. So, sorted data is this in column 6. So, got done make a scatter product the diagram. So, this is now I will explain that. Obviously, it would be normal, would not be normal.

(Refer Slide Time: 15:59)



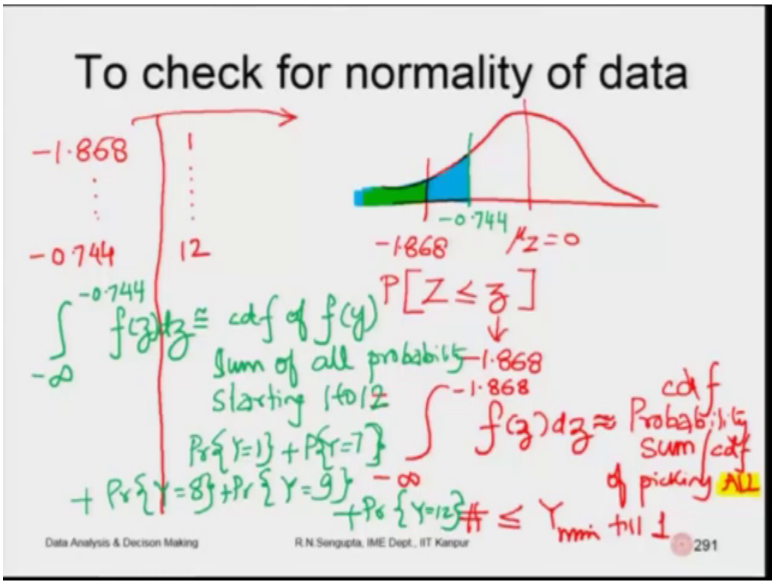
	E	F
2	-1.868241655	1
3	-1.403412636	7
4	-1.128143645	8
5	-0.919135522	9
6	-0.744142742	12
7	-0.589455798	13
8	-0.447767518	22
9	-0.314572292	34
10	-0.186756121	34
11	-0.061931623	44
12	0.061931623	49
13	0.186756121	51
14	0.314572292	56
15	0.447767518	59
16	0.589455798	65
17	0.744142742	65
18	0.919135522	71
19	1.128143645	85
20	1.403412636	90
21	1.868241655	99

So, what we have done is the concentrate on these values I will highlight it make a different color scheme that will be better. So, we will make it as orange make it as light green. So, expand this. So, I would switch on here, this is the one, ok.

Now, it actually what it is doing is this if you concentrate on the orange column starting values minus one point 1.86 still plus 1.86. So, these are the corresponding values on the z table or z distribution with the standard normal whose overall CDF; that means, from minus infinity to that value of minus 1.86 would be equal to the corresponding probability which we are getting when we find out the probability of having number one out of the total number of readings which is there in front of us. So, if I want to find out the corresponding probability of 1, so, for 1, no, I cannot write it here anyway. So, let me use the. So, let me note down these values minus 1.86 let me count it 3 the 3 decimals minus 1.868 and 1.

So, what I am doing is that let me make or tables blank sheet is separately, ok.

(Refer Slide Time: 17:44)



To check for normality of data

-1.868 | 1

⋮ | ⋮

-0.744 | 12

$\int_{-\infty}^{-0.744} f(z)dz \cong$ cdf of $f(y)$

Sum of all probabilits starting 1 to 12

$Pr\{Y=1\} + Pr\{Y=7\}$

$+ Pr\{Y=8\} + Pr\{Y=9\}$

$+ Pr\{Y=12\}$

$P[Z \leq z]$

$\downarrow$

-1.868

$\int_{-\infty}^{-1.868} f(z)dz \approx$ Probability sum / cdf of picking ALL # $\leq Y_{min}$ till 1

Data Analysis & Decison Making

R.N.Sengupta, IME Dept., IIT Kanpur

291

-0.744
-1.868
$\mu_Z = 0$

So, that would better this is the extra slide I am basically inserted. So, let me go it here. So, what was the value was minus 1.86. So, let me make the tables according me and that value was one only considered on these things would become clear total number values for the right hand side was basically 20. So, 1 would have basically a probability that it is being picked up from those 20 numbers. So, obviously, it will be 1 by 20 because there is only 1 1.

Now, for this value of minus 1.686, what you do is that this is the normal distribution this is the mean value which is 0, standard deviation is 1. Consider this minus 1.868 is here. I want to find out the overall area of z starting from total probability starting from minus infinity to that small z which we is I want to find out the probability of this one should be the case where small z is. So, that would be equal to minus infinity to minus 1.868 f of x f of z sorry my mistake f of z dz that will be equal to, let me use a different color.

So, this is here that technically should be equal to the probability, the sum of the total probabilities not the probability cdf probability sum which is the cdf value of picking

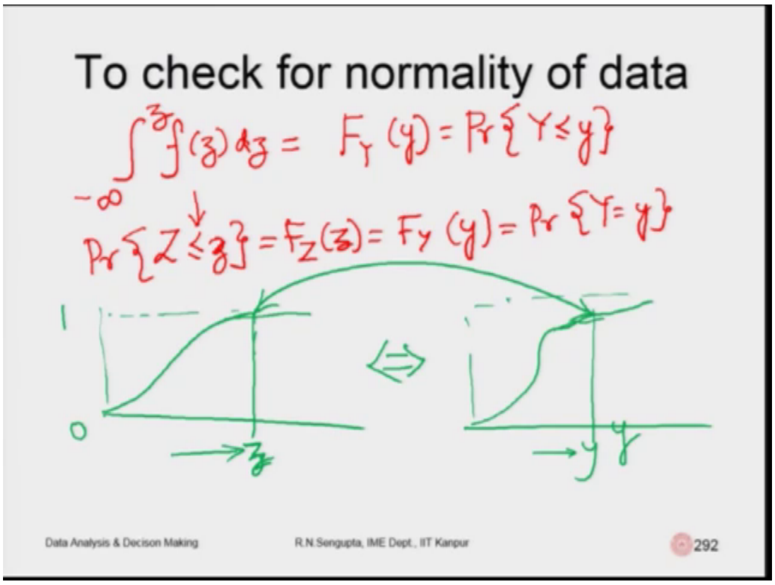
ALL. I am going to highlight this what ALL. All numbers less than equal to; that means, these value so, consider those values I am ranking consider them as Y so, Y minimum till the value which is 1. Now, if I am changing so, this total. So, see this would be technically the cdf value.

Now, if it is changing consider this values here consider the value of minus 0.744 and 12. So, there is 1 2 3 4 5 five values till 12 and the value is minus 0.744. So, let us come here again . So, it will be somewhere minus 0.744, this is 12 there are other values here I am not including them. So, I use a different color you can select the green 1 minus 0.744 is here . So, now, the actual values with the sum of all these was minus infinity to 0 minus 0.744. So, in that case the corresponding equation would be I will using green color. So, in order to differentiate that sum and due to positive place I am I am basically space I am writing it here. So, please bear.

So, is it minus infinity to minus 0.744 f of z dz this is not exactly equal, but it should be cdf value of this f of y whatever the f of y is we do not know. So, that will be sum of all probability starting one to 12. So, I will add up all the probabilities. So, probability of y is equal to 1 plus probability y is equal to what was the next value let me check was 7, 8, 9, 12 ok; 7, 8, 9, 12 probability of 7 plus probability of 8 plus probability of 9 plus probability of 12. So, if add up all the probabilities and that probability which is the cdf and that cdf would be exactly equal to should be exactly equal to the cdf value for the standard normal from minus infinity to the value of minus 0.744.

Now, actually what you are doing is a one to one correspondence and what is that correspondence is I will create another chart another blank sheet and explain I do not want to erase this just pay attention here and I will request that if the focus is there on the slide and it will be much easier for me to explain. So, what we are doing is that trying to find out the cdf values starting from minus infinity to that value of z. So, that z s we have already found out and that if they are exactly equal to the cdf values for all the sum of all the property starting from the least value of the distribution to that value of the value realize value of the distribution we will basically have if there is a one to one correspondence and then the Q-Q plots would be a straight line 45 degree line. I am going to come to that later on.

(Refer Slide Time: 25:23)



To check for normality of data

$$\int_{-\infty}^{z} f(z)\,dz = F_Y(y) = \Pr\{Y \le y\}$$

$$\Pr\{Z \le z\} = F_Z(z) = F_Y(y) = \Pr\{Y = y\}$$

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 292

So, what we are trying to do is, so, this is actually what is the essence. So, what we are doing the summation from minus infinity to some value of z, f of z. This is the standard normal that should be equal to F of y or small y which is probability of capital Y being less than equal to small y. So, this is also equivalent to probability of small z less than equal to capital Z less than small z. So, this is equal to capital of Z small z is equal to capital of Y, this is what we want to ensure probability of Y is equal to y. So, these are the cdf values.

I should have written in a different color wait for the diagram I am going to do that. So, generally this is the cdf value cdf value for the. So, technically is going to infinity. So, this is the z values and this is 0, this is 1. So, that should have a one to one correspondence another distribution which I have plotted see their values I am trying to draw it as nicely as possible cdf values. So, this would basically be Y I have a corresponding value small y. So, the area we check what are this one to one correspondence which I am basically trying to highlight.

And, if you see plot it this one, so, this is not a Q-Q plot. So, if you see that one which you have drawn. So, in this case it basically mode skewed on to one direction, then go straight and then go changes the skewed direction I will come to that this portions later on also. It is going toward I am going a little bit slow please bear with me and once I am going to utilize that for different type of distributions (Refer Time: 28:02) so, Q-Q plots would be discussed in more details at on.

The slides which have written I am not going to delete them, let them be there in the next class again I will basically start discussing that in further details or that it would be easier for you to appreciate what I have done and then try to basically use it your excel sheet. Is very simple, just use this the excel sheet and double verify whatever I have said and the results which will get using exponential distribution or normal distribution in place of the RB distribution which I have plotted would give you good results in order to basically authenticate what I am saying is true.

With this I will end the twenty eighth lecture and continue further discussion about the Q-Q plots in the twenty ninth lecture and the subsequent one.

Have a nice day and thank you very much.