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

Lecture - 42
Introduction to Utility theory

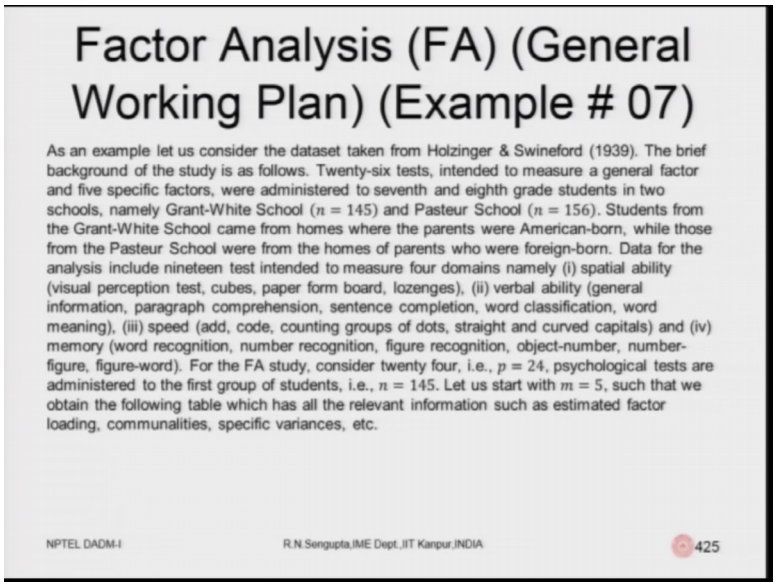
A warm welcome to all of you my friends and students a very good morning, good afternoon, good evening to all of you. And as you know, this is the data analysis and decision making one course under NPTEL, MOOC. And this hold course is for 12 weeks which is 30 hours and we are in the 42nd lecture. Starting we have already completed 8 weeks. So, we are starting in the 9th week. So, ninth week second class and each week we have about 5 lectures, each being for half an hour. And my name is Raghu Nandan Sengupta from the IME department, IIT Kanpur.

So, if you remember, we were discussing in the last class, what is a first one, I give a lot of background, a recap sort of thing for the multivariate analysis concepts, mainly being the first one with the principal component analysis and we are basically doing the factor analysis. So, then the factors, they would be factors depending on P number of variables which are there which may be definitely dependent.

And we take M of them such that we are able to give to the maximum possible extents by the rotation on the factors in such a way that predictability based on the factor plus and error terms would be there. Obviously, the error terms would be kept as low as possible in order to give the so called the maximum efficiency of the factors. But the word efficiency, I am using in a very general sense that you are able to predict, forecast or gives informations from based on the factors.

Let us consider an example, I will read the example decide and give the results rather than going through the actual solution. I will come to that later on when we do the multiple linear regression. If, you remember I did discuss that I will be doing the multiple linear regressions in details and before that they would be another background which I will come most probably by the 42nd or the 43rd class.

(Refer Slide Time: 02:23)



As an example, let us consider the data set taken from Holzinger and Swineford, 1939. The brief background of the studies as follows. Twenty-six tests, intended to measure a general factor and five specific factors, were administrated to the seventh and the eighth grade students in two schools. So, there are two schools, you will not basically have some factors based on which you want to study. So, the study can would have been say for example, background can be whether their parents are educated with their, they have higher degrees or the income bracket from which they belong.

So, whether they have some, some effect, it could have been like whether they have sibling, older sibling, younger sibling, whether they live in a joint family, whether they live in a nuclear family, all this could have been taken. And they are different psychological studies, sociological studies which may be considered when you want to find out some objective criteria based on which you can comment on the factors affecting some output. So, the output could have been the performance of the students. So, output could have been say for example, the general mental balance of the students so, all these that could be considered.

So, let me continue reading it. So, these five specific factors were administrated to seventh and eighth grade students in two schools, namely the Grant White School where the total sample size was 145 and the Pasteur School where the sample size was 156. So, this 145, 156 or whatever the sample size ways from two different schools has would not

basically be the same, it could be different obviously, as it is in this example. It could have been that Grant-White School numbers of observations could have been higher with respect to the Pasteur School and all these things are possible.

Students from the Grant White School came from homes where the parents were American born. So, this is what I, if you remember I mentioned the sociological, the genetic factors could be there from, from which background they are coming, income and all these things are there. Students from Grant White School came from homes where the parents were American born while those from the Pasteur School were from homes of parents who were Foreign born.

Data from analysis includes nineteen tests to in a intended to basically understand four main domain set of information for the students mean namely which are special ability; special ability would be basically the visual perception of trying to understand geometric figures and whether they are able to tell actually as it is.

They would be the concept of verbal ability how they are able to express them in words, whether they can understand the meaning of simple words, understand the meaning of simple spelling, spelling mistake, whether they are able to at least mention that in words what they want or what they are trying to express. So, it could be paragraph comprehension, as I said and related to the logical thought process being expressed in words. They could be comprehension reading, they could be sentence completions, sentence writing, trying to find out the error in the sentence, word classifications, antonym, synonym, meaning of a word. They could be basically speed of reading also.

So, third specific m factors or the main criteria could be speed, speed related to how we can do the addition, subtraction we can count in groups whether you can do your normal activities of trying to keep your plate in the basin, if you are a kid. So, this is basically for the kids, ninth and seventh and eighth standard. Whether you are able to clean the plate as required, whether you want to be able to eat it properly without much wastages so, all these things are noted and how fast you can do it. So, how we can count the group of dots, join the group of dots whether they are a straight line, curved line and so on and so forth.

And the last effect of the factors of the main criteria of concern (Refer Time: 6:49), the study would be memorizing or memory. So, whether you can two word recognizing from

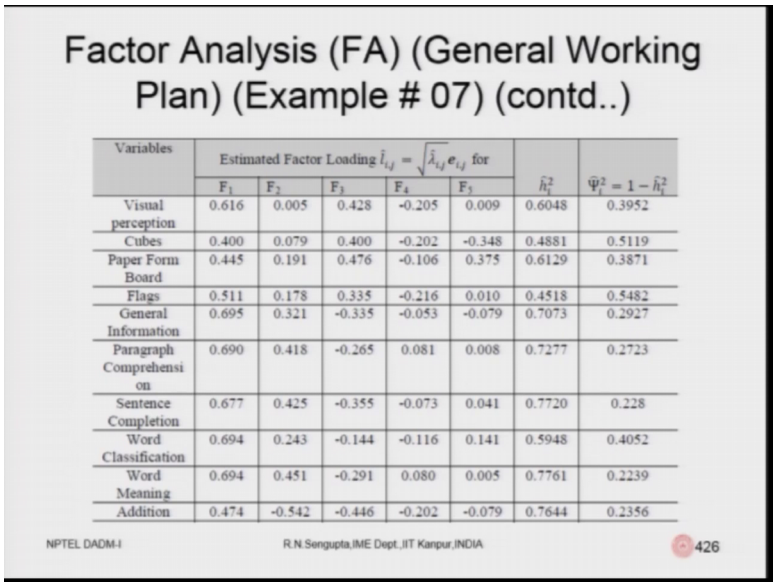
the past sentence, recognition from the past whether you can understand a figure, you can understand see for example, number plates of car, phone numbers to whole names and all these things would be there.

So, can be relate to how you correlate between figures and words, figure written in, in a numeric and the words written in the numeric in the alphabet form whether you can make a one to one corresponds to that. For the factor analysis study consider 24 psychological tests. So, these are the actual test which is x 1 to x p and they may be independent or dependent so, but we will consider they are dependent.

So, this psychological tests are administered to the first group of students which is 145 which is from the Grant White School. So, let us start with m is equal to 5. So, we want to basically find out of these number 24 which is p 1 to find out which 5, in case they give us the best results. Such that one obtains the following table which has all the relevant information such as estimated factor, loading communalities, specific variances, all these things are there. So, what we are doing is we are taking 5 then it could be 7, that can be 9 that can be 15.

So, as an example I am giving 5 such that one can utilize this information to understand what is the factor loading should be there in order to get the maximum amount of information which is required. Given the fact, there are in this case 26 random variables x 1 to x 24, sorry 24 random variables x 1 to x 24 which can basically give us the maximum amount of information from the data.

(Refer Slide Time: 08:42)

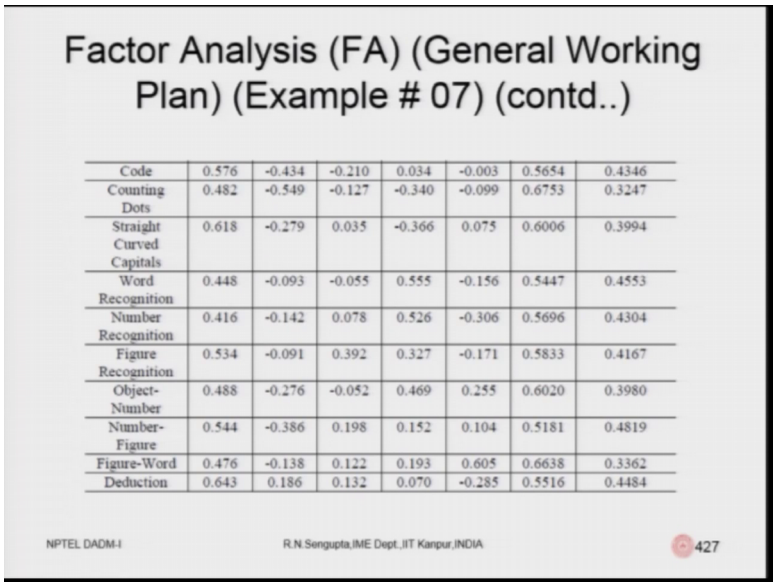


Variables	Estimated Factor Loading $\hat{l}_{i,j} = \sqrt{\hat{\lambda}_{i,j}}\, e_{i,j}$ for					$\hat{h}_i^2$	$\hat{\Psi}_i^2 = 1 - \hat{h}_i^2$
	F_1	F_2	F_3	F_4	F_5		
Visual perception	0.616	0.005	0.428	-0.205	0.009	0.6048	0.3952
Cubes	0.400	0.079	0.400	-0.202	-0.348	0.4881	0.5119
Paper Form Board	0.445	0.191	0.476	-0.106	0.375	0.6129	0.3871
Flags	0.511	0.178	0.335	-0.216	0.010	0.4518	0.5482
General Information	0.695	0.321	-0.335	-0.053	-0.079	0.7073	0.2927
Paragraph Comprehension	0.690	0.418	-0.265	0.081	0.008	0.7277	0.2723
Sentence Completion	0.677	0.425	-0.355	-0.073	0.041	0.7720	0.228
Word Classification	0.694	0.243	-0.144	-0.116	0.141	0.5948	0.4052
Word Meaning	0.694	0.451	-0.291	0.080	0.005	0.7761	0.2239
Addition	0.474	-0.542	-0.446	-0.202	-0.079	0.7644	0.2356

So, these are on the estimated factor loadings. So, if I find out the factors F 1 to F 5 so, for the case of visual perception so, these are the categories which you have already done the visual perception cubes, paper from boards you can formulate you can have general information, paragraph comprehension.

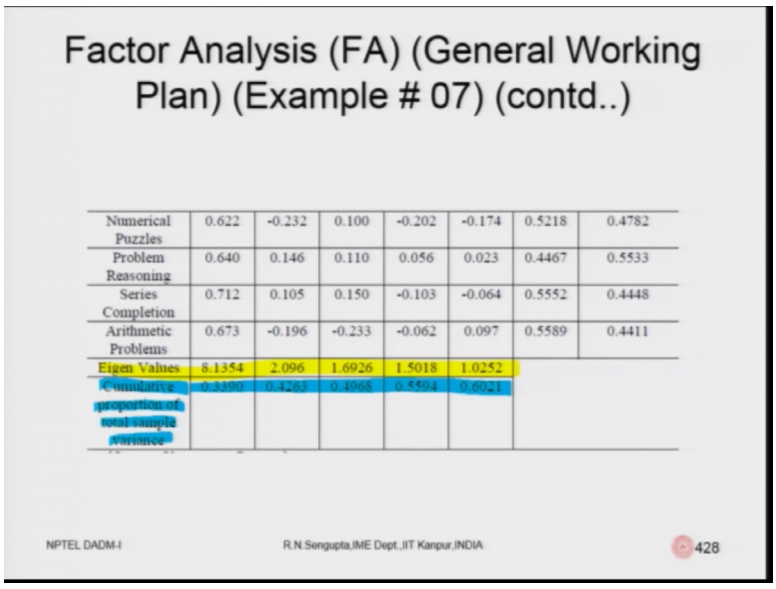
So, these are I am writing them one group of at a time in the sequence. So, in the next set of effects based on what we think are important factors I will not use the word factors, basically are not the important so, called reasons and which are the outputs. So, it is basically paragraph comprehension sentence, comprehension word, classification, word meaning, additions, and all these things. And the factors F 1 to F 5 are given, corresponding to that you have the errors and basically you can pass judgment accordingly.

(Refer Slide Time: 09:36)



So, let us continue. So, in the same way you have the code counting dot, straight curve, capitals word recognition. And again the corresponding factors F 1 to F 5 are given in the second column, third, fourth, fifth, sixth, seventh, fifth, sixth F 1 F 2 F 3, F 4, F 5. And the errors and the corresponding other values are also given the last two column.

(Refer Slide Time: 10:01)

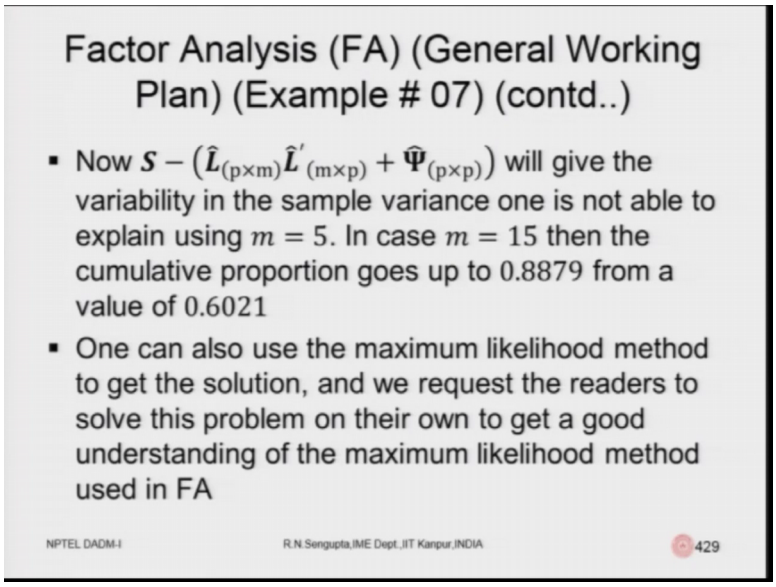


So, cumulative proportions if you want to find out. So, once you do the studies cumulative proportions, the total sample variances and that means, what is the overall

effect, you are able to find out the eigenvalues are given correspondingly. And; obviously, we will see that they are in the decreasing number. So, the first factor would be 8.1354, then 2.096, 1.9626. So, these are the values.

And the cumulative proportions based on which you can predict the variances, the variability appears on color. So, there are cumulative proportions of total sample variances as given as third, I am (Refer Time:11:04) reading the percentage wise 33, 42, 49, 55 and 60 based on the factors F 1 to F 2. So; obviously, you could have the factors of F 1 to F. So, calculations would be done accordingly, your Eigenvalues are would be calculated and then the cumulative proportions the total values would also be found out.

(Refer Slide Time: 11:24)

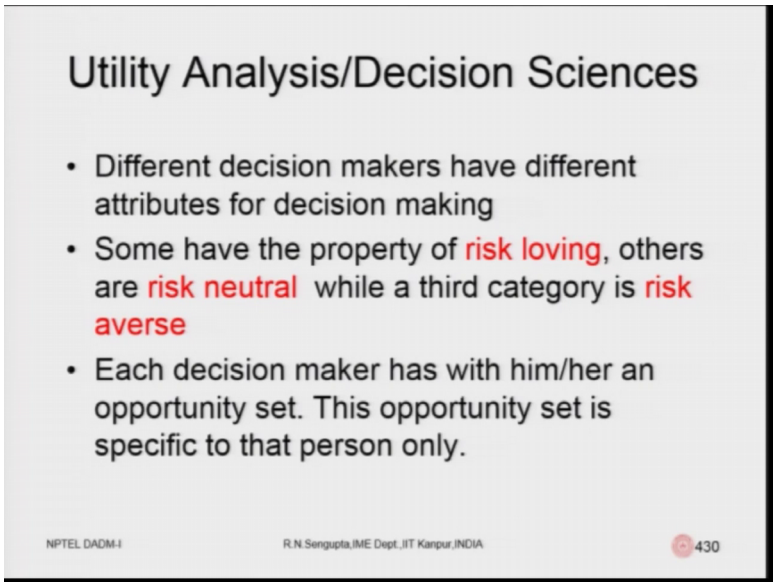


Now, if you want to find out the difference between the so called sample related variance, covariance and the factors. So, that will be, will give the variability in the sample variance one is not able to explain using this m is equal to 5. In case of m is equal to 15, then the cumulative proportions goes up to about 88 percent from a value of 60 percent. So, the values which you is of it was 60 percent from 5, in the case of 15 88, but the reason so; obviously, we will be tempted to use more and more factors. But it will basically depend what is the marginal so called values you are getting as you keep adding one variables at a time, factors at a time.

So, to what level you should go such that you are able to give the prediction to the maximum possible level. So, one can also use maximum likelihood methods to get the

solution and we request the readers may basically they can do some solve the problems on their own to get a good understanding on the maximum likelihood method as it can be utilized and based on that you can compare the factor loading method which we have just discussed.

(Refer Slide Time: 12:31)



Now, with this considering principal component and factor analysis I just change the topic and come into general utility analysis and decision making, again come back to this concept of multivariate statistics later on.

So, this would basically be clubbed as a decision tree or decision sciences area and it will have some consequences how we are trying to analyze the multivariate statistics methods later on. I will give very simple examples which in without going to the details.

So, different decision so, generally when we consider utility analysis or decision sciences. So, we consider the different decision makers have different attributes for decision making attributes are characteristics. So, some of the characteristic can be objective, some of the characteristic can be subjective they can make characteristic based on which we are taking the decision where the numbers are there.

There can be other characteristics in when you are taking a decision where no numbers are there, say for example, I want to buy a car, I am giving an example may not be actually relevant immediately. I want to buy a car then for me the cost factor, resale

value, prices of diesel or petrol whatever it is or maintenance cost all these things would be variables based on which I can take us objective decision, I am using the word objective and subjective not to basically give some colors or some value to the decision. It basically, I mean some that I have some numerical values which I can assign to the objective and the subjective part.

So, when it is something to do with numbers they would be objective part and such that the numbers would be easily communicable with the other person, if he or she is needs to understand why that decision is being taken. In case, if it is basically a subjective one; obviously, in that case rather than objectivity, subjectivity would be more important like say for example, it can be the color of the car, it can be say for example, what a (Refer Time: 14:38) best information set which you have received for the car when you are trying to buy that car. So, I am only considering the problem of buying the car.

It could have been say for example, I want to basically find out how good a bad the school is. So, the school say for example, reputation would very be very important depending on how what the type of plus 10 and 12 results are.

So, in that case I would basically assign some numbers. So, the number of students who are we receive more than 95 percent, number of students who are we receive more than 90, what is the number of students who have been admitted in the professional degrees after class 12, it can be chartered, it can B.com, M.com, then it can be economics, it can be history, mathematics, somebody goes for the engineering one. So, all these things can be considered.

In other can say for example, I may consider the schools. So, how far it is from my residence or what is the locality, is it a good one? Or what is the actual fees based on which I am going to understand that what is the annual fees I have to pay for a kid's education in that school. It can be also say for example, the quality of the faculty members, how good they teach, what whether they have foreign degrees or it can be whether they are willing to listen to the kids or whether it may be say for example, some of the parents things that thinks that a playground is very important. So; obviously, all those facts the f this, I will not use the word Eigenfactor. So, the criteria based on which I am going to take a decision can be both objective and subjective.

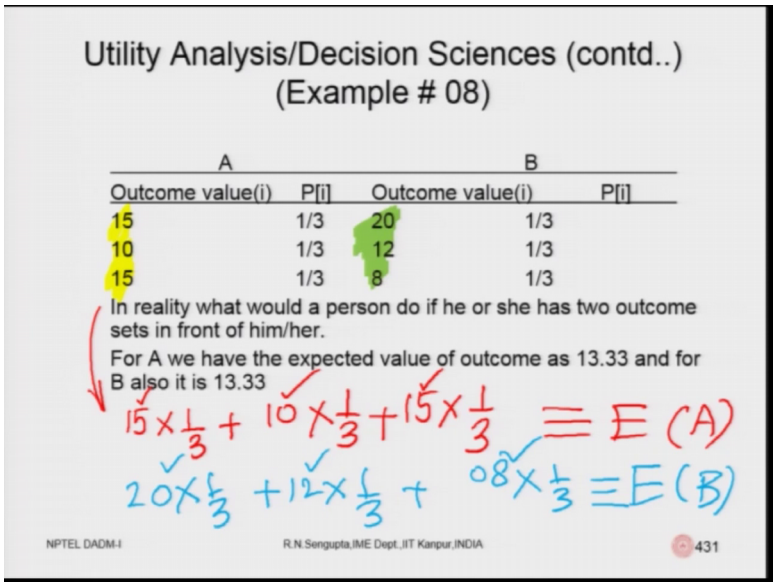
Now, whenever I am taking a decision for decision where so, where monetary outputs are there. So, I am going to consider now monitory outputs with where the numbers can be assigned this concept would be more objective in nature. Now rather than being in the subjectivity in the discussion, we will come to the concept where you analyze a decision making both considering both the objective and the subjective part would be taking up later.

Some have the property, some person set of persons who wants to take a decision has the property where he or she loves the risk. Others can be whether they are indifferent to risk and the third category can be that basically they are risk averse. This one averse means they are willing to run away from the risk. The first set would basically try to go towards the risk, second category would be the set of people who are indifferent.

So, each decision maker has with him or her an opportunity set based on which he or she is going to take the decision and this opportunity set is specific to the person only. So, if the opportunity set for me is say for example, x 1 x 2 x 3, I am giving very simple connotations on the from the vector perspective another person for the same set up of decisions which are in front of him or her the opportunity set based on which he or she is going to take the decision whether following that combinations of x 1 x 2 x 3 or not to willing to take any of the decisions. Obviously, he or she would have a different so called preference number or the opportunity set. So, I am using the word preference opportunity and so on and so forth, inter changeably we will understand it.

So, whenever these cases are there we will not need to basically analyze that there is some utility, some net worth of a decision process which needs to basically be followed such that we know that there is some objectivity in the decision which can basically be converted in to numbers and then decisions can be taken accordingly.

(Refer Slide Time: 18:11)



So, consider this example you have the outcome given in the leftmost column which is 15; 10 and 15. So, these numbers which have put are just theoretical, I will come to the actual usage of the numbers later on, they are just theoretical values the second column gives me the corresponding probability. Now, it technically means for an outcome value of 15, there is a chance that it occurs out of the 100 such occurrences one - third number of the time which is about 33.33 percent numbers of the total 100.

Then if I go to the second column on the second row on the left hand side so, it means that the with the outcome value of 10 again the corresponding probability is one - third and finally, for outcome of 15 the probabilities is one - third. So; obviously, you may be tempted to combine both for outcome A the first stage and the third stage; that means, for which both the outcome value is 15 in both the cases. So, what is and the probabilities are one – third, one - third. So, why did not we combine them? So, answer, I will see is in this way, these values of 15 10 and 15 which you have are the actual.

So, called numeric values based on which you as a third person are seeing that the decision maker is taking. So, the first 15 is say for example, on the amount of so called investment or the so called amount, amount of money required, amount of numeric things which are required in order to facilitate the decision process. But what that value first value of 15 has to me for whom the probability is one - third and similarly for the

third case that when the outcome value is 15 and also the probability is one -third, these two 15s may accrue different connotation for me as an investor as a decision maker.

So, if I am considering the first case of 15, it may mean than the actual conversion based on what my actual initial investment or initial values is 15 that total value of 15 can basically be considered as a decision which would give me a certain net worth which would be definitely be different from the net worth which I am getting from the third decision which means that values may be same, but the net worth for the decision would change from one decision to other that is why I am basically taking, taking them separately.

So, the probabilities are one–third, one–third, one-third. For the second decision maker which is B, the outcome values are 20 12 and 8 then out and the corresponding probabilities are also again one – third, one – third, one - third.

In reality in the question is, what would a person or the decision maker do, if he or she has two outcomes sets in front of him or her? That is the actual so; obviously, these values of 15 10 and 15 with the corresponding values of one – third, one – third, one – third, this is one set. And in the other set, set means not that numerical set, the not that arithmetic one these are the decisions which I have. And for the set B now the second one the outcome values are given as 20 12 and 8 and the corresponding probabilities are one- third, one- third, one- third.

Now, these values for A and for B, these are the values which are apparent to a third party, but when the person is actually taking a decision based on 10, 15, 10, 15, 20, 12 and 8, the actual accrued value to me as a decision maker would be a some sort of functional form based on what is the input which is basically 15 or 10 or 15 in the first case for A and for the second case which is for B, it can be 20, 12 and 8.

Now, when I need to basically find out the values based on which I as a decision maker, I am going to take the decision, it would mean that there is some functional form based on which the input comes as the outcomes which is the first column and the third column and the end result which basically accrues to me would basically depend on what my outlook of so called decision making is and what that outlook I will come to that within few minutes.

So, in reality what would a person do if he or she has two outcome sets in front of him or her? Now, when we basically try to analyze the problem the best solution would be the easiest understanding on that problem would be that you basically find out the expected value. So, the expected value would be given by the case, when for the first case and in the next case for the second one so, 20 into one – third, 12 into one- third 8. So, what I need to find out is so called expected value for B, this is the expected value for A, you combine them and then basically take the decision.

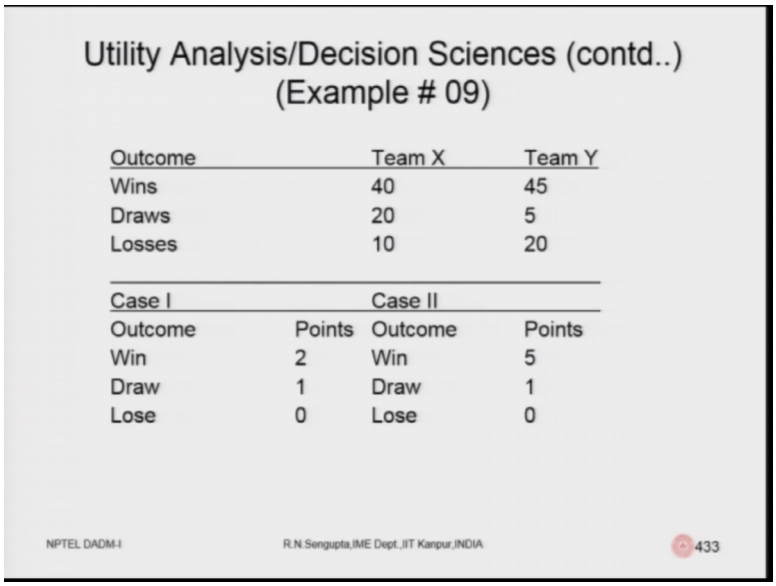
So, the values which are written here 15, 10, 15; 20, 12 and 8, this would basically depend on what my utilities or the net worth based on which I am taking in the decision. So, I will come to that utility concept later on.

Now, consider the same example, but with the different probabilities. The probabilities are for the 15 10 and 15 for the for decision A, the probabilities are not now one- third, one- third, one- third. Now, they are 50 percent, 25 percent, 25 percent. And corresponding to B again it remains that outcome values are 20, 12 and 8 and the corresponding probabilities are one – third, one – third, one - third.

So, in this case if I consider the so, called expected value and the expected value would be 15 into half plus 10 into one - fourth plus 15 into one- fourth, that would with for case A. And for case B, it would basically be 20 into one - third plus 12 into one - third plus 8 into one - third. Again you find out the expected value, but once you find out the expected value you have to basically trying them according to the expected value.

So, now for A, we have the expected value of outcomes given as 13.75 in this case and for B it still remains 13.33. So, in the first case, both the values were coming out with 13.33. Hence, we were indifferent could not take a decision. In this case as now A is expected values 13.75 and B is continues to remain a 13.33. So, our if I main task is to find out the best. So, called outcome will definitely take A into consideration.

(Refer Slide Time: 26:23)



Consider a, a next example which is a different setting I consider a game is being played. And there is team A and team X and team Y. And; obviously, for each in this game, these teams appear for two different tournaments, consider the football 1 the set of teams (Refer Time:26:47) say for example, Santhosh Trophy and the same other team say for example, place some other tournament which is being held.

Now, in that case for the for case one the points accrued for an outcome which is win and draw and loss, the corresponding points are 2 1 0 and when I come compare that for case two the other tournament, the win draw and loss where our values are 5 1 0.

Now, what we need to do is that basically try to combine the decisions. Before I combine the decision, I need to find out whether there we are doing the calculations properly. So, in the first case what we will do for team X and Y, will basically multiply the actual values of the so called values, value means how many such wins are there, how many such loss draws are there, how many such losses are there? And for each corresponding of win draw and loss for case 1 which is the tournament I have signed a point which is 4 2 X would be 14 into 2 plus 20 into 1 plus 10 into 0.

Similarly, when I do it for case 1 for team Y, it will be 45 into 2, 5 into 1 and 20 into 0. So, based on that, I find out the points when I do it for case 2, the second tournament again it will be 40 into 5 now, because the ranking, the point system for second

tournament is different. So, it will 40 into 5 plus 20 into 1 plus 10 into 0. And again, in the similar case for team B, it will be 45 into 5, 5 into 1 and 20 into 0.

So, I will with this, I will end this 42nd lecture continue in discussion the utilities in the 43rd, –44th and 45th. And slowly build up that how we can use the decision concepts for the, along with the utility theory for analyzing different type of decisions under the case where there are very specific examples in multivariate statistical methods.

Thank you very much and have a nice day.