

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

**Lecture – 11**  
**Loss Function**

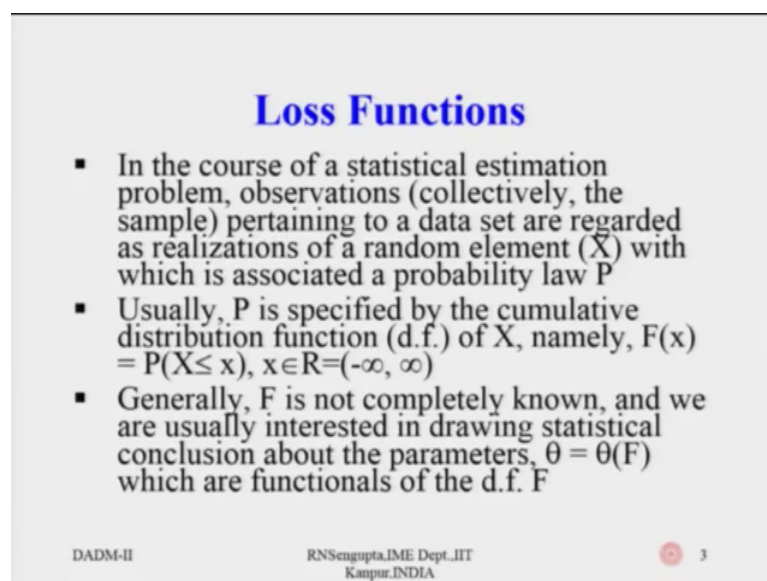
Welcome back my dear friends, a very good morning, good afternoon and good evening to all of you and hope all of you are fine. A very happy new year to all of you considering that I am recording this after 2019 started. So, I everything hope everything is fine at your respective end.

Now, as you know this is the DADM - II course, Data Analysis and Decision Making – II and we are going to start the 3rd week lecture and this whole course is for total number of hours is 30; that means, each week we have 5 classes or lectures each lecture being for half an hour. So, in totality we have 60 lectures and we are going to we have completed 10, we are going to start the 11th lecture. And my name is R N Sengupta from the IME Department, IIT Kanpur.

Now, if you remember we were discussing UTD functions, then concept of safety first principle, the concept that how different bounds, the Chebyshev's bound, Marco bounds could be utilized and based on that how problems could be formulated as in a simple optimization problems, where you want to have a look at different concepts of. These are not the directed use of utility function but they are we have you could use optimization, and when you are using optimization; obviously, the concept of utility function would be utilized time and again.

So, in today's lecture we will discuss something to do with not a different concept a similar concept depending on how we can utilize it as loss functions. And why loss functions are important I will come to that. And we will try to utilize that in this, I will give you examples and then try to utilize it later on how loss functions could be utilized. And remember that these loss functions would be important when you are trying to analyze different decision making where nonparametric decisions are important. So, let us consider what we mean by loss functions?

(Refer Slide Time: 02:43)



**Loss Functions**

- In the course of a statistical estimation problem, observations (collectively, the sample) pertaining to a data set are regarded as realizations of a random element ( $\bar{X}$ ) with which is associated a probability law  $P$
- Usually,  $P$  is specified by the cumulative distribution function (d.f.) of  $X$ , namely,  $F(x) = P(X \leq x)$ ,  $x \in R = (-\infty, \infty)$
- Generally,  $F$  is not completely known, and we are usually interested in drawing statistical conclusion about the parameters,  $\theta = \theta(F)$  which are functionals of the d.f.  $F$

DADM-II RNSengupta,IME Dept.,IIT Kanpur,INDIA 3

So, in course of any statistical study, any type of estimation, any type of forecasting, any type of prediction we have some set of observations. Now, this set of observations are taken from the pool; that means, you have the population, population I mean the whole set of observations which are applicable from which you can take the data and you take a certain observations. Now, any for any statistical study we ok, by the way we did discuss few of these things in the DADM – I, but I will again repeat it, it would be needed that is why. Why it will be needed I will come to that later. So, please have the patience and please be here with me.

Now, for any statistical observations or studies or estimation, prediction, forecasting, we as I said that we pickup set of observations those are smaller in number. And our main aim is basically to utilize those set of observations to draw some meaningful conclusions. So, as it says in the course of meaningful conclusion about the population because you have a small sample and the sample characteristics should be such which you are going to study because based on the data you want to predict or forecast something about the population characteristics. It can be the mean, median, standard deviation, the value of the ratio of the mean to median, what is skewness, kurtosis whatever it is.

So, in course of statistical estimation problem observations collectively known as the sample pertaining to a data set are regarded as realization of the random variable. So, each observation when you pick up, they are realized value of the random variable and

with which is associated a probability law or a distribution. So, we will basically mean mentioned in (Refer Time: 04:32) basically it is small f of x or capital F of x depending on how you are trying to explain that. Small f of x would basically be for the pmf or the pdf and capital F of x would be the cdf function which is the cumulative distribution function.

Usually P is specified by the cumulative distribution function as I just mention namely F of x and its basically the summation of all the properties from the minimum value of X to that value particular value of X based on which we are trying to find out the cumulative distribution function.

Generally, F is not completely known and because there are parameters in the distribution. So, they can be the shape, scale and location parameter which we denote by for the pdf or the pmf; shape, scale and location, alpha, beta, gamma. And these alpha, beta, gamma, are for the population when you pick up a sample we may not know that we want to estimate this is the general task. So, generally F is not completely known and we are usually interested in drawing statistical conclusion about the parameters, where theta which is basically function of the distribution function.

(Refer Slide Time: 05:49)

### Loss Functions

- In a parametric model, the assumed functional form of F may involve some unknown algebraic constant(s), which are interpreted as the parameters, e.g., in the normal d.f., the algebraic constants are themselves the mean ( $\mu$ ) and the variance ( $\sigma^2$ )
- The objective in point estimation is to utilize the information in the given set  $(X_1, \dots, X_n)$  of sample observations (random variables) to choose a suitable statistic  $T_n = T(X_1, X_2, \dots, X_n)$  such that  $T_n$  estimates  $\theta^n$  (parameter) in a meaningful way

$\lim_{n \rightarrow \infty} P\{ |T_n - \theta| \leq \epsilon \} \rightarrow 1$

DADM-II
RNSengupta, IIT Kanpur, INDIA
4

So, our main (Refer Time: 05:49) is this. Now, in a parametric model the assumed functional form of a F may involve some unknown algebraic constants which are interpreted on the parameters as I said. Example in the normal distribution you know the

mean which we generally denote by the mean  $\mu$  and standard deviation  $\sigma$  they are generally the parameters  $\mu$  and  $\sigma$ . Then say for example, you have in the exponential distribution  $\lambda$  and  $\theta$ , in the Poisson distribution  $\theta$  the binomial distribution  $n$  and  $p$ . So, there are different (Refer Time: 06:18) for each distribution which are as I said, again I am mentioning from the population are unknown.

So, when you pick up a certain observation which is mentioned here. So, you pick up an observation and they realize (Refer Time: 06:33), they are random variables which is  $X_1$  to  $X_n$ . So, actually when you pick them up you have and they realize they become small  $X_1$  to small  $X_n$ . So, they are known to you. These are the sample observations random variables to choose a suitable and based on these random variables  $X_1$  to  $X_n$ , we choose the statistic. Statistic is basically the characteristics which you are getting from the sample which is  $T_n$ , suffix  $n$  is basically means that the number of observations. So, it is a functional form of  $X_1$  to  $X_n$ . So,  $T_n$  is basically function of  $X_1$  to  $X_n$ .

So, it can be if you are trying to find out the mean it can be the weighted mean, simple mean, exponential mean whatever it is. Such that  $T_n$  estimates, so what we are interested is basically to find out that how does the  $T_n$  which you have found out basically gives us some information about the  $\theta$  which is the parameter which you are working on, which you want to find out from the population.

Now, whenever you are picking up the set of observations and when you found out  $T_n$ , our main concern is that that  $T_n$  should be as close as possible to  $\theta$ . So, if you remember we have discussed two important properties when we were doing DADM - I. So, one was basically unbiasedness; that means, the expected value of  $T_n$  in the long run too should be equal to  $\theta$  and one was consistency in the sense that as the sample size  $n$  increases the probability of the difference between the  $T_n$  value and  $\theta$ ; that means, the bound slowly tends to 0 as  $n$  tends to infinity.

So, if I write it down technically the first would be I will write the first one erase it and then write the second one considering the space is limited here in this slide. So, the first one with a expected value which unbiasedness which you have already consider. There are proofs for that but I am not going to go into that.

So, let us erase it. So, then we have the unbia; the consistency which means limit  $n$ ,  $n$  is the sample size tends to infinity, probability of  $T_n$  minus  $\theta$  ok, being less than equal

to yes I should basically specify it less than equal to some epsilon. Epsilon is the very small value tends to 1. That means, the difference always becomes as slow low as possible. So, it is basically is the variance in the long run would be decreasing. So, if this properties are met; obviously, we are happy with this the this parameter over this estimate  $T_n$ .

Now, the issue is the two things whether it is actually possible to find out  $T_n$ . So, answer is that in many of the cases  $T_n$  may not be, may not be not able to find out  $T_n$  even if we consider these two properties or even if we are we ignore this two properties the functional form of  $T_n$  may not be possible to at least delineate or give an expression.

(Refer Slide Time: 10:03)

**Loss Functions**

- Imposing *consistency* and *unbiasedness* does not always lead to a unique estimator of  $\theta$
- A good idea is to locate an optimal estimator within the class of consistent (and possibly, unbiased) estimators
- One idea is to choose a nonnegative metric  $L(T_n, \theta)$  defined for all  $\theta$ , where  $\theta$  varies over (the parameter space) while  $T_n$  varies over (the sample space, which is usually a subset of  $R^n$ , the n-dimensional Euclidean space)

DADM-II RNSengupta,IME Dept.,IIT Kanpur,INDIA 5

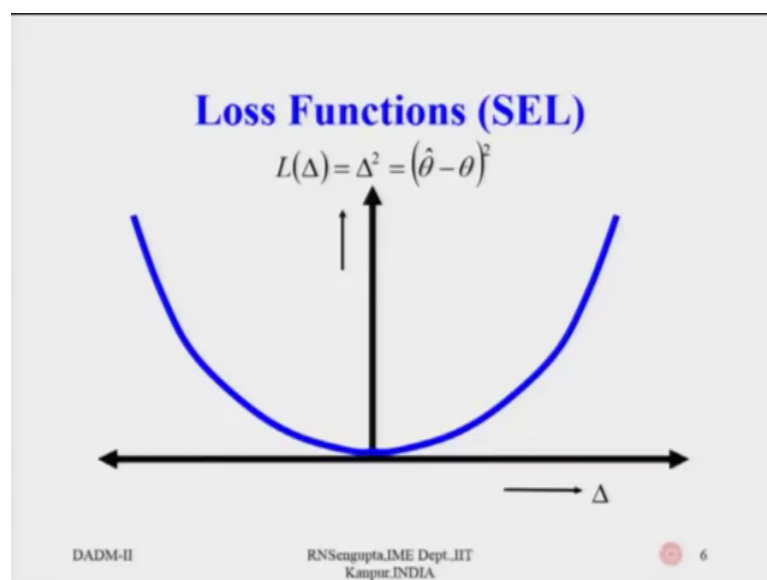
So, coming to this imposing consistency which I mentioned and unbiasedness does not always lead to a unique estimate of theta. So, some may be consistent, some may be unbiased and all these combinations. A very good idea would be to locate an optimal estimator within the class of this consistent and possible unbiased estimator which will do our work or which will give our result. So, consistency would more be more important and if they are unbiased and obviously, well and good, so, will basically concentrate on consistency.

One idea would be to choose a non-negative matrix. So, we want to basically measure it and that measure what we do is known as the loss function which is defined for all values of theta and all values of  $T_n$  which we find out why, where  $T_n$  varies over the sample

space and it is a subspace of the real number and depending on the number of parameters which are there.

So, what we are interested to find out is basically find out  $T_n$  which is a functional form  $X_1$  to  $X_n$  which is the realized value such that the functional form of a loss; loss means some type of negative gain we are going to have in choosing  $T_n$  such that there is a difference between  $T_n$  and  $\theta$  in their actual value that will give us some loss, some negative value. So, you want to basically have a functional form of that loss and try to basic understand that how that loss can basically be optimized or reduce; obviously, we will try to reduce it because if  $T_n$  is closer to  $\theta$  well and good, it is actually what we want if  $T_n$  is definitely far away from  $\theta$  that is actually do we do not want.

(Refer Slide Time: 11:45)



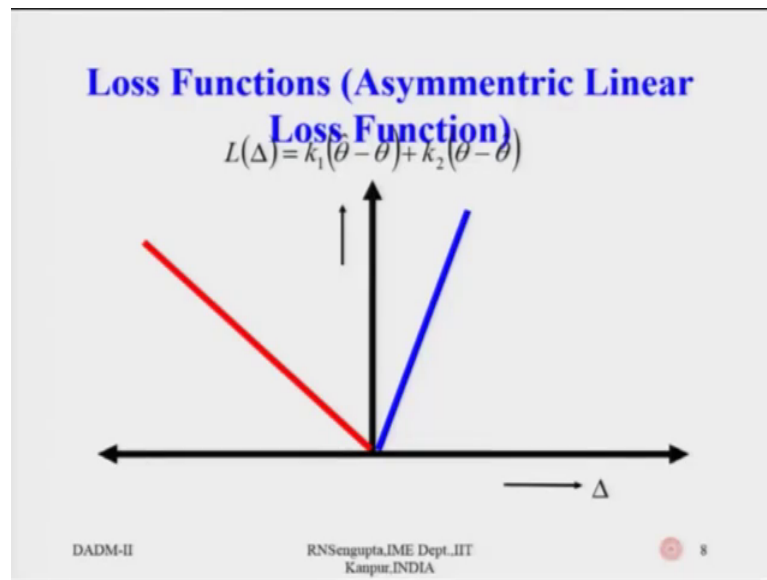
So, few of the loss functions which you may have seen in DADM – I, but I will again I mention it. So, one is basically the quadratic loss function or the squared error loss function. So, in this case you see that if I basically find out and draw a delineate or mark delta. Delta is basically this, so which is basically the difference between  $\theta$  hat and  $\theta$ .  $\theta$  hat is basically  $T$  and which is the estimate we of  $\theta$  we are going to find out from the sample. So, the difference of  $\theta$  hat and  $\theta$  is basically delineated along the x axis and the loss function is basically delineated or drawn along the y axis. So, in this case this is a quadratic loss function, so hence the loss function which we will draw along the y axis would be  $\theta$  hat minus  $\theta$  whole square.

So, all the values which are positive would be drawn on the right hand side which considering  $\hat{\theta}$  is greater than  $\theta$  and all the values which are negative will be drawn on the left hand side, on to the origin which is  $\hat{\theta} - \theta$  is negative. But, obviously, as the loss function is quadratic, so if  $\hat{\theta} - \theta$  is plus 2 and  $\hat{\theta} - \theta$  is minus 2 the square value becomes 4. So, both on the right hand side and left hand side of the graph that is in the first quadrant and second quadrant the height of that function  $f$  of  $x$  which is basically  $L$  of  $\delta$  is basically of equal quantum.

Now, the advantage of this loss function is that if you try to minimize this loss function it gives us some information of trying to basically minimize the variance. Because what is variance? Variance is expected value of  $\theta - \hat{\theta}$  and try to find out the expected value of that which is basically the dispersion which we are talking about the variance. And if you minimize that; obviously, it will lead to the fact that minimizing the loss function conceding quadratic loss function also leads to the minimum variance. Another, which is good and obviously, there are good results for that theoretically is very nice but; obviously, practically it will not be nice as I will try to give some examples later on within few minutes.

So, another loss function which could be utilized would basically the mod loss function. So, which is linear loss function in the sense that both positive and negative loss functions are of equal width. So, hence and the difference of  $\hat{\theta} - \theta$  if it is positive it is given a positive value, not a square value in on the first quadrant and if  $\theta - \hat{\theta}$  is negative then also it is given a mod value which is also same quantum on the second quadrant be. So, obviously, it will be a 45 degrees line both in the in the first quadrant and the second quadrant but in the first quadrant is positive and the second quadrant is negative but you are trying to delineate is accordingly.

(Refer Slide Time: 14:36)



Now, consider this, the loss functions which I just discussed was the mod loss functions which is linear loss functions. Now, consider the positive part is weighted by a factor which can be more than 1 or less than 1. So, in that case in this example which we see in front of us which is in the slide the value of which you are giving to the weight which is when it is on the positive quadrant which is the first quadrant considering theta minus theta hat is possible positive is more than the weight which you are going to give it to in the when it is in the second quadrant.

That means, for any positive deviation we are going to give a linear weightages; obviously, linear functional form would be there but the weightages would be much more. While if it is in the negative direction in the second quadrant we will give a weightages which will be less than the first quadrant, which means that if the difference is say for example, plus 2 and in other case the difference is minus 2 we will multiply plus 2 by unit. Which is say for example, 1.5 but in the case when we are getting a minus 2 we will only multiply it with 1 or less than 1 which means that we are penalizing the penalization of the loss for positive 1 is higher and similarly for the loss for the negative 1 is lower.

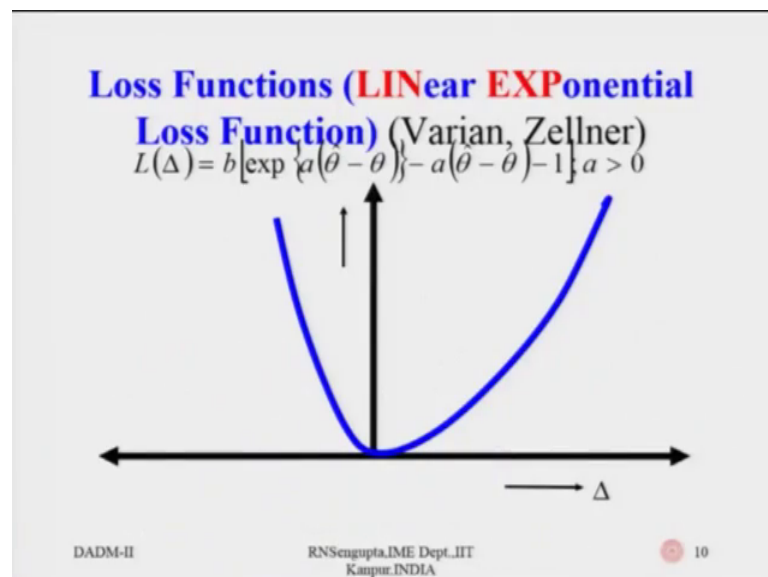
Then obviously, you will think then why not switch it; that means, we give higher weightages for the losses which are in the second quadrant and give lower losses comparatively lower losses which are for the losses which are in the first quadrant. So



obviously, it will happen that in that case the loss would look like this when the weightages given to  $\theta - \hat{\theta}$  minus  $\theta$  which is positive which is the blue one. The weightages which you are giving for that is less than and that is what we are drawing in the first quadrant is less than the value or the weightages which you are giving in for  $\theta - \hat{\theta}$  which is negative which is then pink one in second quadrant. So, obviously, we are trying to unequally penalize the losses for the positive and the negative losses but they are linear in nature.

Now, technically this linear part has a in issue because if you see the line at the origin so obviously, there the concepts of continuity and differentiability are an issue, I am not going to go into the details about that. So, all variant in the 1960s gave a very nice loss function which is known as the linear exponential loss function and that was mainly in the concept of trying to find out in the economics perspective. So, this asymmetric or linear non-linear loss function was basically known as LINEX loss. So, the part LINEX the first part a line basically means, the LIN basically means the linear part and EX basically means the exponential part.

(Refer Slide Time: 17:33)



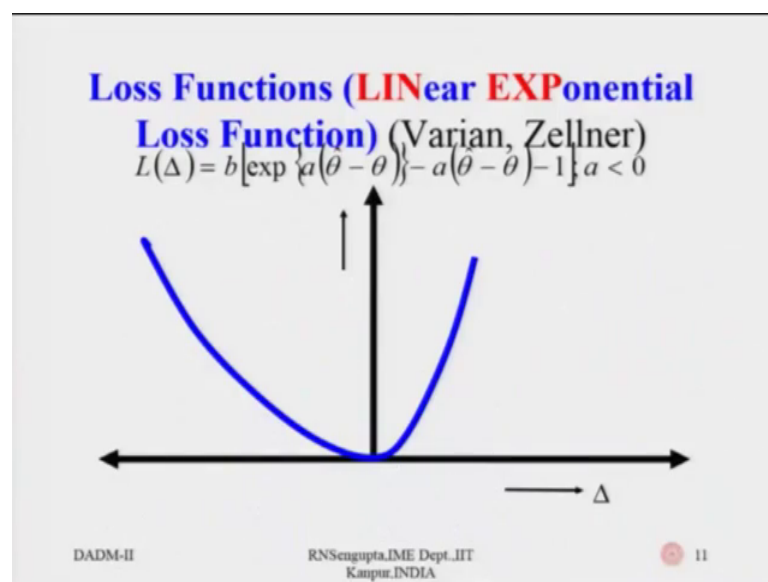
So, if you if you see the slide the loss function is linear exponential a LINEX loss function. This p should not be there (Refer Time: 17:41), so it should be small. So, LINEX loss function is a part, if you check this loss function forget for the time being the parameter b, only concentrate in the equation form which is given inside the square

bracket. So, you have basically exponential  $a$  into  $\theta$  minus  $\theta$  hat that is the first term which is exponential part for the LINEX loss and the second part which is linear is as minus  $a$  into  $\theta$  minus  $\theta$  hat minus  $\theta$  and there is a minus 1. Now, why this loss function is useful let us go slowly.

If you consider the this the loss function for the part and as this shown in the graph, if  $a$  is positive remember that means, the parameter is positive, then as  $\theta$  minus  $\theta$  hat increases which is the first quadrant the weightages on the overall loss which is coming out from the exponential part slowly starts dominating the linear part. Hence, the part in the first quadrant is much more penalize than in the case when it is in second quadrant. That means, when  $a$  is positive and you are basically trying to find out  $\theta$  minus  $\theta$  hat which is negative. That means, overestimation is more penalized than the underestimation in the case for the LINEX loss when  $a$  is positive.

Now, let us consider the picture when  $a$  is negative and  $a$  still you basically want to delineate or draw the LINEX loss function considering  $a$  is negative and both for the first quadrant and second quadrant.

(Refer Slide Time: 19:18)



Now, here the graph looks like this. In the sense when  $a$  is negative when you are trying to basically take  $\theta$  hat minus  $\theta$ . So, in that case the exponential part will dominate the linear part on the second quadrant. And when again in the same place when  $\theta$  hat minus  $\theta$  is positive and  $a$  is negative, then in that case linear part which will dominate

the explanation part because that negative part would bring that ratio down as fast as possible. So, in this case you will have the exponential dominating in the second quadrant for  $\hat{\theta} - \theta$  being negative and  $\hat{\theta} + \theta$  being positive the linear part will dominate, which means that again we have a case where it is an asymmetric loss function but underestimation would be more dominated than the their over estimation.

Now, consider this loss function that why it is important? Let us basically consider that  $\hat{\theta} - \theta$  is very very small this is in an around the value of origin which is 0. So, if you basically expand the exponential form, so the first term would be 1, the second term would basically be plus  $\hat{\theta} - \theta$  divided by factorial 1. The second term is basically again it will be plus a square into  $\hat{\theta} - \theta$  whole square by factorial 2 and the terms will continue. So, if you consider the first term 1 and minus 1 they will cancel out as given here, I am not writing it out please listen to me carefully you will understand.

So, basically if you expand again I am repeating, if you expand the exponential part the first term is 1, so 1 and minus 1 which is already there in the LINEX loss function cancels out. The second half part for the expanding of the exponential part would be into  $\hat{\theta} - \theta$ . In the numerator divided by 1 factorial in the denominator, so again that cancel us cancels out with the second term which is given in the LINEX loss function. The third part in the expansion of exponential part is plus a square into  $\hat{\theta} - \theta$  whole square which is in a numerator divided by factorial 2 and it will continue.

Now, consider  $\hat{\theta} - \theta$  is a very small number. So, obviously, square is a smaller number and cube and other half powers basically are quite smaller number. Say for example,  $\hat{\theta} - \theta$  is 10 to the power minus 2, then  $\hat{\theta} - \theta$  whole square will become 10 to the power minus 4,  $\hat{\theta} - \theta$  cube becomes 10 to the power minus 6 and so on and so forth. So, if we ignore the cubic term the fourth part term, the fifth part term for the expansion and obviously, 1 1 has already cancelled plus a  $\hat{\theta} - \theta$  into  $\hat{\theta} - \theta$  and minus a into  $\hat{\theta} - \theta$  already cancelled.

So, you are only left with the second the third term which is the quaternary term. So, you will see for very small values of the difference between  $\hat{\theta} - \theta$  the LINEX

loss function almost exactly belongs to, is almost the quadratic loss function. Which means a special property of the LINEX loss is that it basically penalizes both overestimation it can also penalize underestimation depending on how you basically frame the problem. And in the special cases when  $\theta - \hat{\theta}$  is almost closer to 0, the LINEX loss basically can be replaced by the quadratic loss function and all the properties of the quadratic loss functions can be utilized.

Now, we will basically discuss three simple examples. In case, in order to basically make you understand that how these values of  $a$  being positive, being negative and  $\theta - \hat{\theta}$  whether overestimation important or underestimation important or whether we depending on the problem, we will take the situation with where overestimation is more important or underestimated is more important.

(Refer Slide Time: 23:12)

**LINear EXPonential Loss Function**  
(Example # 01)

Consider a company plans to launch a new product, say a refrigerator in the consumer market. Also suppose that similar products from different manufacturers already exist in the market. Then the company is expected to give some warranty for the particular product, i.e., the refrigerator, to its customers in order to sell the product. Now if the value of this warranty is more than the average time of failure for the product, then the aforesaid mentioned company needs to replace the damaged products it sells, or face litigation charges. On the other if the warranty period is less than the average failure time of similar products available in the market, then the company loses the market share to its rivals, as naturally, customers are willing to buy the refrigerator from the competitors who assure a higher warranty period. Under such a situation it is definitely advisable to estimate the warranty life time using an asymmetric loss. What values of  $a$  one should use would then depend on the level of importance our company places on overestimation versus underestimation, i.e., the cost of litigation versus the cost of a loss in the market share of the company

DADM-II RNSengupta,IME Dept.,IIT Kanpur,INDIA 12

So, consider a company is going to launch a new product and the company is launching a new product in a sense that there are other competitors also, I will read the statement which is given in the slide but let me first explain it.

So, here the company is launching a product considering that is a refrigerator or a TV or a fridge and AC whatever it is. Consider for the time being the product which you are launching or you are launching the market, that warranty life is say for example, given as 6 months. Now, there are competitors and 6 months you have estimated for yourself from the marketing study.

Now, consider two scenarios; scenario one the competitors have a similar type of product but their actual warranty life is less which is basically 4 months. So, in this case what happens? 6 months you basically give a product, so in that case when you when the product is basically floated in the market people will be more tempted to buy your product because you are giving a warranty of 6 months so obviously, the initial sales for your product is much higher. But remember, the actual product which is already there in the market similar type of product which is been floated by which has been floated by a competitor that 4 month is the actual warranty life which should have been but somehow either due to your erroneous calculations or due to your marketing strategy you gave a warranty life of 6 months.

Now, initially the product was sold quite well but generally the products would be start failing in and around fourth month so obviously and you have already given a warranty life of 6 months so obviously, you have to pay some penalty to the customers and basically repair their products from the money from your own pocket. Which means the initial positive thing has basically will, now be wiped out by the negative thing which means that this over estimation and underestimation has to be analyzed in such a way that the market which you captured in the initial stages whether it is loss due to the loss of sales as it happens later on.

During, after the fourth month when you basically find out that people are basically not been satisfied as the product has started failing in and around the fourth month in spite of the fact that you have basically given a warranty life 6 months which was wrongly calculated. So obviously, you can understand which is more, initial sales was better or whether it was basically better to give you the warranty life of 6 months.

Now, consider the other way. You actually the warranty life is 8 months and you have given a warranty level 6 months. So, now, remember one thing 6 minus 4 and 6 minus 8 these both things are plus and minus, minus 2 and plus 2. So, if you basically consider from the point of view of the quaternary loss function it will be 4 4 in both the cases.

Now, let me come back to this example. So, it is basically 8 months and you have given a 6 months. So, initially the products sold by you would be much less because people will be more attracted for the warranty life which is 8 months but as they find out the their products which they have brought from the your competitor starts failing so obviously,

people will be more attracted to your products later on because they find out the warranty life is actually what it should be which is 6 months and slowly you gain your market.

So, initially you had a loss in the sense the market share was less. Now, you will slowly gain the market share. So, in the first example scenario one is that initially you gained a big market share but slowly you started losing the market share and the confidence of the customers because many of the product which you sold in the market had to be repaired because of the warranty life which you stated as 6 months is not actually 6 months is basically 4 months.

So, in this case overestimation or underestimation is more penalize or less penalize you have to basically analyze this problem basically state the condition based on which you will basically frame this problem. So, now, with this let me read the problem.

Consider a company plans to launch a new product say a refrigerator in the consumer market. Also suppose the similar products from different manufacturing already exist in the market then the company is expected to give some warranty for the particular product, that is the refrigerator to his customers in order to order to sell the product. Now, if the value of this warranty is more than the average time of the failure for the product then the aforesaid mentioned company needs to replace the damaged product it sells or face litigation charges. Obviously, if you give a warranty life which is less than what it is actually should be or more than then actually obviously, in one of the cases they would be there would be some litigations.

On the other hand if the warranty period is less than the average failure time of a particular product similar products in the market, then the customer company loses the market share to its rivals as naturally customers are willing to buy the refrigerator from the competitors who assure a higher warranty period as I mention. Under such a situation it is definitely advisable to estimate the warranty lifetime using an asymmetric loss function whether you want to penalize or less penalize or more penalize will depend on this situation what you are facing.

What values of  $a$ , that is the parameter which I said  $a$  being more positive that is  $\theta$  hat minus  $\theta$  would basically be penalize more; and in the case when  $a$  is negative then  $\theta$  hat minus  $\theta$  which is negative would be penalized more. Because exponential

part will dominate in the first quadrant the first case when  $a$  is positive and exponential part will dominate in the second quadrant for the case when  $a$  is negative.

So, what values of  $a$  one should use would then depend on the level of importance our company places on the overestimation versus underestimation, that is the cost of the litigation versus the cost over loss in the market share of the company has to be analyzed.

So, with this obviously, there would be more other examples coming up for overestimation or underestimation specifically, but with this I will end this 11th class and continue with the discussion later on for the from the 12th class. And try to basically give you an example where the loss functions could be utilized and later on see how we will try to basically merge it with the utility function and asymmetric loss and try to solve many of the nonparametric problems which are there to be discussed for this course.

Have a nice day and thank you very much.