

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

Lecture – 01
Introduction

Very good morning, good afternoon and good evening my dear friends. Welcome to this three part lecture series, three part in the sense that, it will have three different courses titled data analysis and decision making as you can see on the screen, in the slide.

And this would be when all of you have gone through this the general syllabus, which was very brief as mentioned from my side. So, it will basically have three big components one is the statistics part, one is the operation research part and one is basically the use of other nonparametric techniques, and computing part. So, this decision analysis and decision making, which we can make as say as DADM 1 would basically be the statistics 1, which will be floating. And we will go basically may pay special emphasis on multivariate statistics, but obviously it needs some background.

So, for the background we will cover few sessions, few lectures for the univariate statistics and for that I will go slowly through data analysis, how what do we mean by probability then discrete distribution, and continuous distribution a concept of expected value and so on and so forth. This course of DADM 1, DADM 2 and DADM 3 would basically be under the NPTEL MOOC, and they would be run back to back in three semesters or sessions.

The first one is DADM 1 which will only focus on statistics. And obviously, the linkage between DADM 1, DADM 2 and DADM 3 would be such that people can utilize the concept so called data analytics and data analysis in a much better way trying to utilize how things or the concepts more from the practical point of view and less from the theoretical point of view that would be our focus can be utilized for different areas of problems it can meet in engineering, can be it in, in marketing, can be in sociology, can be in say for example, quantitative finance can be in operation research and so on and so forth.

So, also to give a very brief background about myself; I am Raghu Nandan Sengupta from the IME department at IIT Kanpur. And this would be the first course as I say that I will be covering DADM and continuing with 2 and 3 so, this is 1.

Before I go through the syllabus you should remember for all the courses, which I have taught like TQM 1, TQM 2, project management, and initially the quantity finance, they were of 20 hours lecture, so 20 hours basically half an hour each lecture for one week, it will be 5 courses, 5 lectures rather than course. And then, there would be an assignment, but this these 3 back to back courses, which I would basically say they are a one big envelope of set of basket, which consists of three different topics would be 30 hours each.

So, 30 hours each would basically be 60 different lectures, each lecture would be of half an hour. And again the same thing would be repeated, and for each week, we will have five lectures of half an hour each. So, we will continue for more than 8 weeks what we have done for the 20 hours one.

So, the first I will go very briefly through the syllabus of DADM 1, then cover few of the textbooks. So, this let me tell you these textbooks are not exhaustive, there are a different type of very classic books in the market, but we will try to basically give some sample of the textbooks, which one can use, and definitely utilize the problem solving, you utilize them from the concept of theory to understand that how DADM 1 can be understood you know in a general sense and where you can apply.

(Refer Slide Time: 04:18)

Syllabus (Data Analysis & Decision Making)

- Elementary probability theory
- Conditional probability
- Bayesian concepts
- Discrete and continuous random variables
- Generating functions
- Central Limit Theorem and uses
- Functions of random variables
- Basic Jointly distributed random variables
- Sampling theory and sampling distributions
- Method for statistical inference

Data Analysis & Decision Making R.N. Sengupta, IIM Dept., IIT Kanpur 2

So, we will consider very briefly the concept of elementary probability theory. We will consider the concept of conditional probability and what does conditional probability mean. We will go through the Bayesian concepts, very briefly and to the point.

Then we will consider the concept of discrete and continuous random variables, what we mean by random variables, and what the concept of random variables probability, theory, chance, relative frequency, what they mean I will cover that also. Then we will go briefly into generating functions and how generating functions give us some concept about, the different type of moments of the distribution. So, the moments as I, as I mentioned it would be the first moment would, would be something to do with the mean on the central tendency, it can be median, it can be mode also, then will consider the concept of dispersion, standard deviation, variances, then we will go to the higher moment third one, fourth one, it would will be skewness kurtosis.

And you will also see that how these moments can basically be obtained from the generating functions and what the concepts in utilization for the first, second, third, fourth and higher moment would be. We will consider the central limit theorem and some of the limit theorems, how they can be utilized, but maybe if we are not able to cover them from the practical sense, but I will definitely try to give you a flavor from the

theoretical, and the conceptual sense. Then we will consider different functions of random variables, how they can have implications in our, in our discussion.

Then we will go to the basic joint distribution random variables for the discrete case, for the continuous case and we will keep it limited to the discrete case, and that for the continuous case we will keep it limited to the Bivariate normal distribution. Then we will go to the concept of sampling theory and why sampling theory is important.

And as we will discussing we will also again repeat, what will be obviously going through beforehand. The concept of sample, sample, space, population and, and all this related concept and we will basically (Refer Time: 06:24) mainly emphasis on the three main sampling distribution, which will be all coming from the normal distribution would be the chi square distribution, would be the f distribution, would be the t distribution. And we also see that how the normal distribution, the normal deviate chi square, t distribution and the f distribution they can be utilized in different type of problems and how we can use the tables to understand them.

Then we will consider the method of statistical inference. The concept of how the concept of point distribution can be utilized, how the concept of interval estimation can be utilized, how the concept of hypothesis testing can be utilized, and obviously, we lay more stress on hypothesis testing. And when we are doing these three topics, which are under the ambit of statistical inference we will cover that what we mean by consistency, what we mean by unbiasedness, and what the significance is with respect to the sampling distribution and the parameters, and their estimates, and what implications they have.

(Refer Slide Time: 07:26)

The slide is titled "Syllabus (Data Analysis & Decision Making)". It contains a bulleted list of topics: Theory of point estimation and estimation of parameters, Theory of interval estimation, Theory of hypotheses testing, Descriptive and deductive statistics, Linear and multiple linear regression, Analysis of variance, and Introduction to statistical Packages, e.g., MATLAB. At the bottom left, it says "Data Analysis & Decision Making". At the bottom center, it says "R.N. Sengupta, IME Dept., IIT Kanpur". At the bottom right, there is a red circle with a white number "3".

We will go to the theory of point estimation, estimation of parameters and as I mentioned what is the concept of unbiasedness and consistency. And as I also mentioned with theory of n^2 an interval estimation we will be utilized for different values of probability for different p values. The hypothesis testing will be considered and we will consider in the conceptual sense, what we mean by null hypothesis, what we mean by alternative hypothesis, which is h_0 and h_a and respectively.

We will consider very briefly the descriptive and deductive statistics, which are utilized. We will consider the linear and multiple linear regression in though brief, and consider what are the assumptions how they can be utilized. We will consider the analysis of area variance or ANOVA and obviously, some statistical packages of MAT lab and all these things. We will come to the statistical packages later on please have the patience.

(Refer Slide Time: 08:20)

Syllabus (Data Analysis & Decision Making)

- Introduction to Multivariate Analysis
- Multinomial, Multivariate Normal, Multivariate t, Wishart and other Distributions
- Multivariate Extreme Valued Distributions
- Copula Theory
- MANOVA, MANCOVA, etc.
- Multivariate Statistical Methods like
 - Conjoint Analysis
 - Cluster Analysis
 - Multiple Discriminant Analysis
 - Multidimensional Scaling
 - Factor Analysis, etc.

Data Analysis & Decision Making R.N.Sengupta, IIM Dept., IIT Kanpur 4

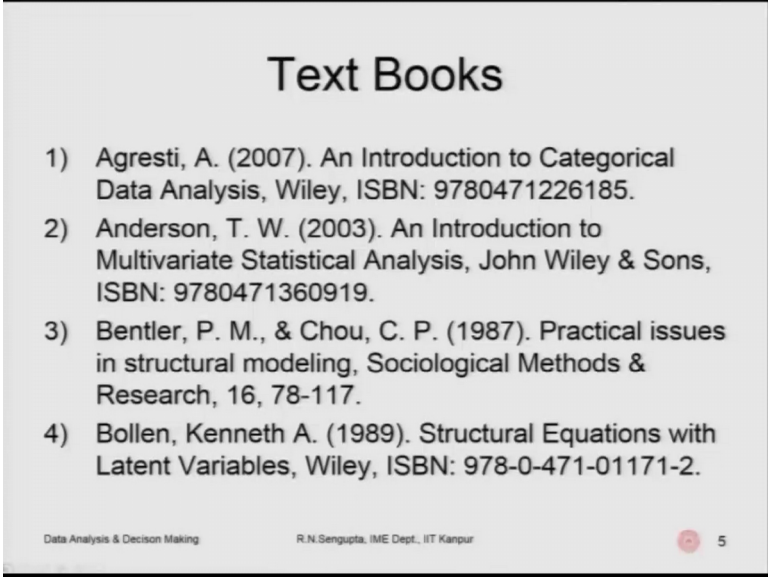
To further continue we will consider the concept of different type of multivariate time statistical analysis, we will consider the multinomial distribution, the multivariate normal distribution, the multivariate t distribution and the corresponding statistics for that the viz distribution, and other distribution of the multivariate, case would be considered and what are the implications with not going to the theory as I am repeating again it will be more from the practical sense. We will consider the multivariate extreme value distribution EVDS of type 1, type 2, and where they are used will leave more emphasis on their on usage.

We will consider the copula theory, and how multivariate distribution copula theory can are being utilized. We will consider the multivariate analysis of variance MANOVA, multivariate analysis of covariance, which in MANCOVA. We will also consider later on at the last part of this course the different type of multivariate statistical methods like conjoint analysis, cluster analysis, multiple discriminant analysis, multi triple, multi dimensional scaling so obvious, factor analysis, structural equation modelling hence when else both and how they can be utilized for different above multivariate statistical studies.

Now, as I mentioned there are 1001 books in the market some are classic books, so obviously, it would be not possible to cover this concepts in many of the cases, but I will try to give you as far as possible their references for the classic books and this classic

books have been the cornerstone based on in which many of the courses for masters, bachelors and Ph.D, courses for different statistical courses are designed. But I will be very brief and to the point and, and definitely now going to the high end. So, some of the books so, obviously, there are again I am saying they are not exhaustive I have a list of about 12, 13 books.

(Refer Slide Time: 10:19)



Text Books

- 1) Agresti, A. (2007). An Introduction to Categorical Data Analysis, Wiley, ISBN: 9780471226185.
- 2) Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, ISBN: 9780471360919.
- 3) Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural modeling, Sociological Methods & Research, 16, 78-117.
- 4) Bollen, Kenneth A. (1989). Structural Equations with Latent Variables, Wiley, ISBN: 978-0-471-01171-2.

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 5

So, we will consider the book of an Introduction to Categorical and Data Analysis by Agrest Agresti 2007. Anderson, which is a classic book in an Introduction to Multivariate Statistical Analysis, but obviously, Indian versions are not available. We will consider Bentler, Chou the Practical issues of structural modeling from the Sociological perspective. We will consider the Structural Equation and equations with Latent Variables book by Bollen which is also a Wiley publication.

(Refer Slide Time: 10:46)

Text Books

- 5) Chatfield, C. and Collins, A. J. (1980). Introduction to Multivariate Analysis, Chapman & Hall, ISBN: 978-0-412-16030-1.
- 6) Duran, B. S. and Odell, P. L. (1974). Cluster Analysis: A Survey, Springer-Verlag, New York, ISBN 978-3-642-46309-9.
- 7) Freund, J. E., (2012). Mathematical Statistics; Prentice Hall of India, ISBN: 8120307844.
- 8) Gnanadesikan, R. (2011). Methods for Statistical Data Analysis of Multivariate Observations, John Wiley & Sons, ISBN: 0471161195.

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 6

The other books which will be considered for different type of multivariate statistics would be a Chatfeild book about multivariate analysis. The Duran book which is a Survey of Cluster Analysis Technique which I use, a very good book for the initial statistical concept would be the Freund book, but obviously this sequence of the books are named as per their the surname of the Author so, it is not divided into areas where I am going to consider. We will consider the methods of Statistical Data Analysis of Multivariate Observations by Gnanadesikan, so this will also be from the mutilator statistical point of view.

(Refer Slide Time: 11:27)

Text Books

- 9) Härdle, W. K. and Simar, L. (2007). Applied Multivariate Statistical Analysis, Springer-Verlag, ISBN: 9783540722434.
- 10) Johnson, R. A. and Wichern, D. W. (2002). Applied Multivariate Statistical Analysis, Pearson Education, ISBN: 8178086867.
- 11) Kotz, S. and Nadarajah, S. (2004). Multivariate Distributions and Their Applications, Cambridge University Press, ISBN: 0521826543.
- 12) Render, B. Stair. Jr., R. M. Hanna, M. E., and Badri, T. N. (2008). Quantitative Analysis for Management; Pearson Publication, ISBN: 9788131723739.
- 13) Seber, G. A. F. (2004). Multivariate Observations, John Wiley & Sons, ISBN: 9780471691211..

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 7

A very good book in Applied Multivariate Statistical Techniques and Their Applications is Hardle, and Simer. And another one which is the tenth one, which is Johnson and Wichern which is the classic book the old it is definitely used. Multivariate Distributions and the application by Kotz and Nadarajah is a classic book from Cambridge.

We will also consider the Render Stair Hanna's from Quantitative Analysis for Management, which is also to do with univariate statistical techniques. And GAF Seber, Multivariate Observation book which is also old book would be another one for the multivariate one.

So, if you see these 13 books, they main stress is basically on the area of multivariate statistics and obviously, that is the main focus of these books. So, we will try to wrap up the in initial introduction for univariate statistics about hypothesis testing as I said then about point estimation, discrete distribution, continuous distribution, expected value, moments as fast as possible.

(Refer Slide Time: 12:38)

Software/Language

- 1) MATLAB <<http://www.mathworks.com/>> . One can find sever based MATLAB at <https://www.iitk.ac.in/ccnew/>
- 2) R <<https://www.r-project.org/>>
- 3) SPSS <https://www.spss.co.in/>
- 4) SAS <https://www.sas.com/en_in/home.html>

Data Analysis & Decision Making R.N. Sengupta, IIME Dept., IIT Kanpur 8

Some of the software languages which are very heavily used in statistics would be the R. So, R is a free software anybody can download it and use it. We have the SPSS, we have the SAS, and another very good tool is the MAT lab, which would be utilized to do both R and mat lab would be utilized to solve the problems from the point of view of multivariate statistical and univariate statistics book.

(Refer Slide Time: 13:03)

Examples

- 1) News paper vendor: wants to maximize profit.
- 2) Production manager: wants to minimize waiting time of jobs on machines and thus reduce inventory and costs.
- 3) Marketing manager: wants to recommend the best changes for a product (which is being sold in the market) so that it will do well.

Data Analysis & Decision Making R.N. Sengupta, IIM Dept., IIT Kanpur 9

So, let us consider an example so, consider the news paper vendor. So, he wants to maximize this profit so, what is the problem, the problem is the news vendor buys papers daily in the morning and he sells them. So, consider he buys them at 2 rupees whether it is Times of India, Hindu states and telegraph whatever it is. And he sells them daily at say for example, let us consider one price at 5 rupees so obviously for each paper if it is sold, he or she makes a profit of 3 rupees.

But consider that if he buys 10 different papers, 8 are sold so which means the 2 are not sold. So, obviously, if 2 are not sold he cannot sell them at more than the cost price he had to sell them as, as papers, as just newspapers which can be utilized for other work. So, obviously, he sells them at a loss.

So, initially he brought them at some amount, sold them at a higher price if there is a market; for those amount, which are not sold he had to sell them at a lesser price so, obviously, the question arrives and comes to our mind is that what is the optimal number of papers he should buy in order to basically maximize his profit. So, obviously, the question is, which is unknown to us is that, what is the number of papers he should buy, and what is the demand of the papers daily.

Next, consider of a example a production manager. He wants to minimize the waiting time of jobs on machines and thus reduce inventory and cost. So, jobs are arriving. They are being processed, and they are being basically packaged, and shipped onto the

customers end. Now, obviously, depending on the demand of the customers, demand on the products, what is the cost price of the products, what is the availability of the products, so, obviously, he will keep an inventory. But, too high an inventory of; obviously, would mean that he has to basically incur an inventory cost and that would basically entail both depreciation both security prices and so on and so forth space constraints.

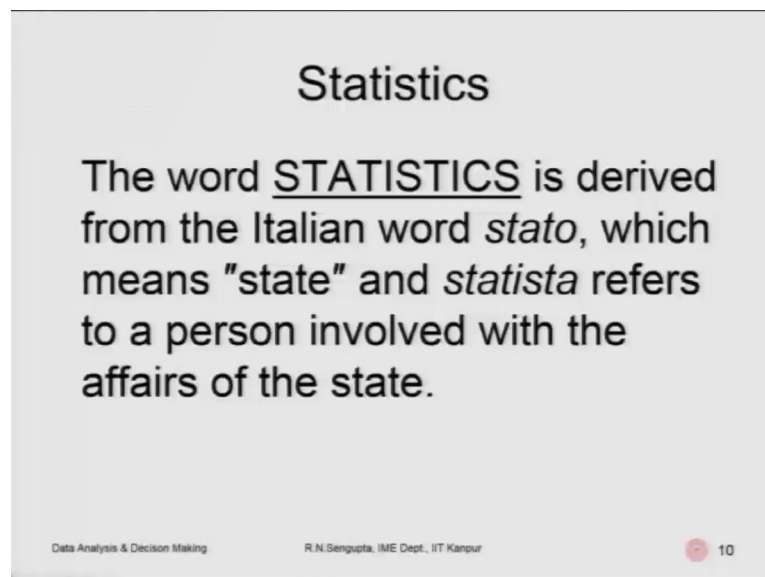
So, his main question would be the vendor, the production manager would ask himself is that what is the demand of the customers, based on that he will basically have the optimum number of products stored in the inventory such that he can work them on the machines and sell them in the market without much of an inventory storages as that is able to reduce, the inventory storage. Plus he is able to reduce the lock in money, which is being he spending, he or she is spending on the inventory and such that cost are minimized and he is able to reduce his inventory.

Let us consider a third example a marketing manager wants to recommend, the best chances of a product which is to be sold in the market, so that it will do well. So, obviously, he does not know the demand, he should know the demand like in the next the second example which I gave for the production manager, he should know about the new one and the customers in the same way the, the marketing manager should know the demand of the customer. And we will basically order products raw materials which you want to basically order and work on them or basically order some products he wants to sell in the market.

Consider he is the marketing manager of one zone which is south of India and he wants to basically have some white goods like Fridge, T V and then washing machines, he want to basically store in his in his distributor end at his factories end, where he wants to ship them to the retailers and from where they will be sold. So, depending on the demand of the customers he would like to basically keep the minimum inventory as well as try to maximize his profit such that whenever a demand comes, he is able to meet that. And if there is no demand; obviously, technically and theoretically he should not have any products, because that would basically clog his money. And, but obviously, if he is he keeps less so if a customer d 1 comes he is not able to satisfy that it is basically a loss sale for him.

So, in this three examples which you saw that will always see that there is some concept or some quantity or some variable, which is unknown, which is not known to you beforehand that would be basically random. And our main task is basically to pay some importance or, or, or note that how the demand, how the supply, how the inventory costs of these things can be considered in order to basically maximize the profit or minimize the loss of the cost. Now, all these things are basically dealt in though and the ambit of the subject of statistics.

(Refer Slide Time: 17:28)



So, what we mean by statistics. The word statistics is derived from the Italian word *stato* which means state and *statista* basically refers to a person, who is involved with the affairs of the states based on which he passes his or her judgment basically that is the concept of statistic, but considering the statistic has come a long way strategy is basically now in encompasses different type of tools, different our mathematical techniques, different type of concepts.

And obviously computing has also become a very important factors such that people can utilize numbers, which are not deterministic, which are random and basically you can understand that randomness of these, these variables and make some judgments based on which some decision can be passed. So, decision means all the three decisions like the vendor boy, the the person who is in the production manager, the marketing manager

basically they can understand what is the number, so that mean they can optimize on their objective function whatever the objective function is.

(Refer Slide Time: 18:29)

Statistics

- Now a days, **STATISTICS** (in a plural sense) is the study of qualitative and quantitative data from our surrounding, be it environment or any system so as to draw meaningful conclusions about the environment or system.
- It also means (in the singular sense) the body of methods that are meant for treatment of such data

Data Analysis & Decision Making R.N Sengupta, IIM Dept., IIT Kharpur 11

Now a days, statistics which basically is in a plural sense is the study of the qualitative as well as the quantitative data from our surrounding, be its environment whatever the environment, where we are operating or any system where we are operating so as to draw meaningful conclusions about the environment of the system and such that we can give some information about the system based on which we are working.

So, in the first case environment is the papers are uncertain how you will sell you do not know, demand you do not know you, you basically have to find out what is the distribution, the demand and to based on that you will make your decision. In the next time an example is basically, what is the demand of the customers based on that you will have the inventory. In third example, it is the demand of the customers based on its which the marketing manager would order the products which can be sold in the market.

The word statistics also means in the singular sense. So, the first one was the plural, in the pluralistic sense, plural sense. In this in the sense in this second example is in the singular sense, it means the body of methods that are meant for treatment of such data such that we can find out some meaningful, meaning or meaningful words based on which we can do some mathematical analysis in order to study them.

(Refer Slide Time: 19:41)

**Main steps in the study of
Statistics**

- Method of collection of data
(primary or secondary)
- Scrutiny of data
- Presentation of data (non
frequency data, frequency
data)

Data Analysis & Decision Making R.N. Sengupta, IIM Dept., IIT Kanpur 12

So, the main steps in the study of statistics are. So, there would be a method of collection of data it can be primary or secondary. So, if it is primary, we will go basically go into the into the environment collect the data. And basically analyze the data such that you have to basically pass on some meaningful judgment about the environment which you are going to study.

In the secondary data, it will be the data which you are collecting from a from a from the primary source like the government has the data you collect it or the census is the data you collect it, financial stock market has the data you collect it, rainfall is a data you collect it, humidity is a data you collated. So, based on that you try to study the environment study the, the actual topic, which you want to analyze and basically pass some meaningful judgment.

Next what obviously, would be whether it is primary or secondary you will basically scrutinize the data. So, you will try to scrutinize the data what whether they are missing data or whether the raw data has some errors or whether they are spurious data, whether they are the data have been collected in the in the random fashion or they have, they are there is have been some effect from my environment, so obviously, you will try to analyze them and make it as clean as possible for our studies.

Then obviously, in the next step would come is basically the presentation of the data it can be non frequency type or frequency type based on which you will try to basically

portray to the set of persons, who want to utilize the data in a best possible of manners such that information can be gathered or gleaned as soon as possible, or as easy as possible, or as fast as possible.

(Refer Slide Time: 21:13)

Main steps in the study of Statistics

- Analysis of data through statistical models/methods
- Conclusions from results thus obtained
- Modification of statistical models/methods depending on results obtained

Data Analysis & Decision Making R.N.Sengupta, IIME Dept., IIT Kanpur 13

The next steps after the presentation would be basically the analysis of the data through different statistical methods techniques which I definitely mentioned when we are going through the syllabus. It can be concept of probability, cost of inference techniques, can be hypothesis testing can be multivariate statistical methods whatever it is.

Once you analyze the data you want to basically conclude. So, whatever the analysis has been you want to conclude and give some answers. So, the next step would be conclusions from the results thus obtained so based on which some decision can be taken.

And in the later on steps you will try to basically modify or a modification of the statistical models, and methods should be done depending on what are the results are obtained, whether the results which are obtained theoretically and matching with the practical results. If they are not obviously, you will try to analyze what are the assumptions based on which you have done the modeling, what is that the scrutiny have done for the data whether if the raw data is, absolutely free of error whether that has been done.

So, if that has not been done obviously, the models would not be exact such that the discrepancy in the results of the models which you are trying to predict or give would basically be quite high with respect to the practical data or practical answer, which you are getting. So, obviously, that they would be feedback to based on which the results which you get from the theoretical sense and the practical sense, you try to match them and basically try to then change your model in order to basically suit and find out the assumptions based on which you are working are the exact such that they give near optimal or actually as close as possible results to the practical sense.

(Refer Slide Time: 22:50)

Descriptive Statistics

Presentation of data

- Non frequency data
- Frequency data

Data Analysis & Decision Making R.N.Sengupta, IIME Dept., IIT Kanpur 14

So, the descriptive statistics which mentioned was basically the presentation of the data can be done as non frequency data and frequency data. So, obviously, we know the frequencies the numbers times of occurrences; and non frequency data would be that we are not using the concept of number of occurrences in order to present the data.

(Refer Slide Time: 23:09)

Non frequency data
Time series or Historical data

Consider the case where the representation of the values of one or more variables like population of India, price of petroleum etc., may be given for different periods of time. For instance we may be interested in knowing the population change over time or the change of production of petroleum over time.

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

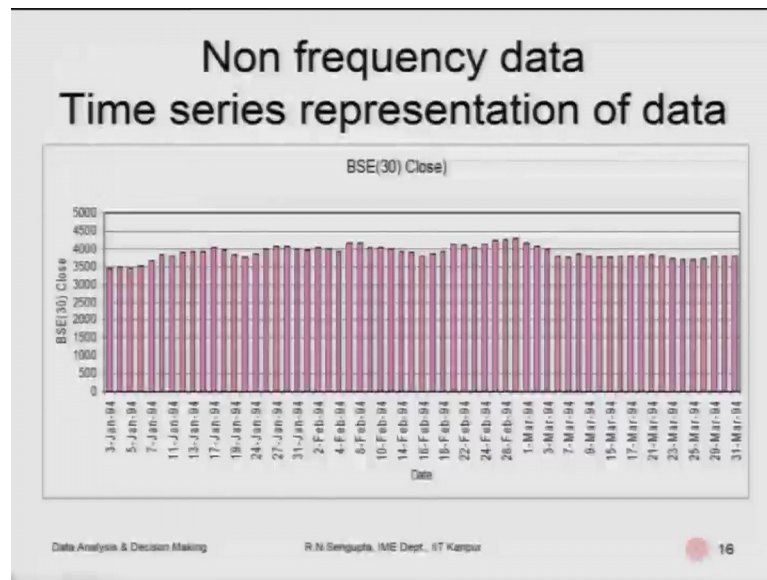
15

So, non frequency data say for example, can be time series or historical data. Consider the case where the present representation of the values of one or more variables like population of India, price of petroleum, price of mica, price of gold, the total the output of grains which is they have, which is there for any particular state say for example, the state of Odisha. So, what is the rice production of Odisha for the last 10 years so, those would basically be non frequency data or on the time series scale.

For instance we may be interested in knowing the population change over time or the change of production of petroleum over time, or the change of production of paddy over time for Odisha or maybe say for example, migration numbers happening between states. So, all these thing, things can basically be the data which can be presented in a non frequency format.

So, let us consider this simple example and they would known by any calculation as such we I am just presenting the data in order to make you understand that how the data can be presented to you.

(Refer Slide Time: 24:13)



So, non frequency data say for example time series representation of the data can be for the Bombay stock exchange the index which consists of 30 stocks. The closing price which you are considering starting from third of January 1994 to 31st March 1994, we have the data as given. So, the BSE 30 closed values are given on the y-axis starting from 0 to 5000; and along the x-axis we have basically the times.

So, the histogram which you have basically it can be drawn in other ways also, it gives you how the prices if you follow the marker how is wearing is the basically the prices, how the BSE 30 piece indices back prices change starting from almost the first of January 1994 till the end of March 1994.

(Refer Slide Time: 25:02)

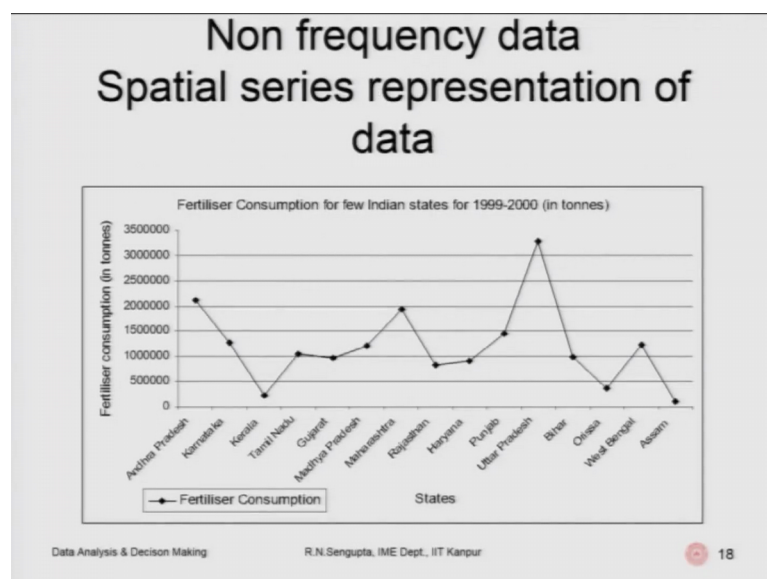
Non frequency data Spatial series data

It may be that the values of one or more variables are given for different individuals in a group for the same period of time. But instead of considering the group as such we may be more interested in studying the way the values of the variable(s) change from individual to individual in that group.

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 17

Non frequency data can be a special series type also it may be that the values of one or more variables are given for different individuals in a group for the same period of time. So, obviously, we in the time frame is same, but they are given for different individuals, you have to basically collate and collecting and present them in such a data that they make meaningful sense. But instead of considering the group as such we may be more interested in studying, the way the values of the variables, change from individual to individual in that group and such that we can find out what is the percentage change for that particular variable between groups which you are going to study.

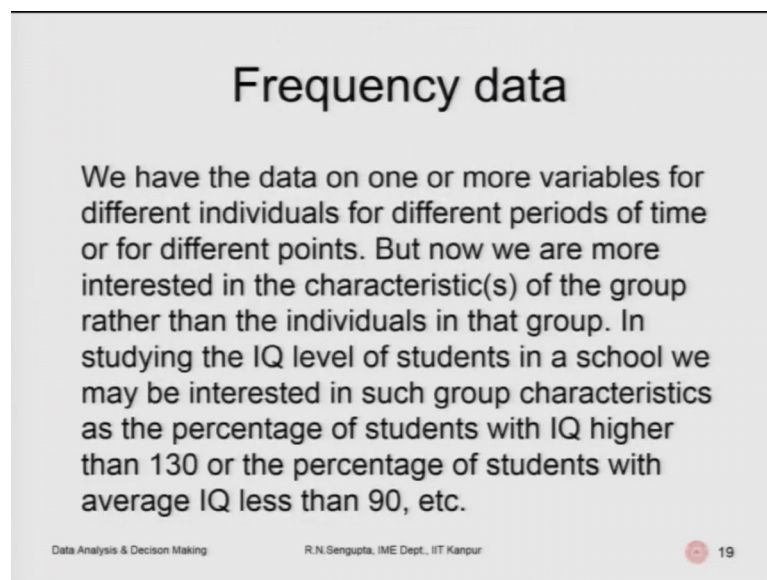
(Refer Slide Time: 25:38)



So, say for example, this is the non frequency data for the spatial series representation of the data where we have the fertilizer consumption for the states starting from Andhra Pradesh, Karnataka, Kerala, Tamil Nadu, till Assam for the year 1999 to 2000 in met in metric tons.

Fertilizer consumption is given in along the y-axis, where the minimum value is 0 maximum value is given as 35 lakhs. And along the x axis you have basically not the time is basically the each and every individual quant, the individuals which are there, which are the states, which is Andhra Pradesh till Assam. And if you see the variation will give you what is the actual consumption of that particular product which is fertilizer for all the different states in the year 1999 and 2000.

(Refer Slide Time: 26:25)



Frequency data

We have the data on one or more variables for different individuals for different periods of time or for different points. But now we are more interested in the characteristic(s) of the group rather than the individuals in that group. In studying the IQ level of students in a school we may be interested in such group characteristics as the percentage of students with IQ higher than 130 or the percentage of students with average IQ less than 90, etc.

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 19

Now, we will come to a few very simple examples of frequency data. So, we have the data on one arm on wave variables for different individuals for different periods of time or for different points. But, now we are more interested in the characteristics of the group rather than the on the individuals in that group.

In studying the say for example, you want to study the I Q level of students in a school we may be interested in such group characteristics as may be relevant which is the percentage of the students with I Q or integer coefficient quotient higher than 130 or the percentage of students with average I Q less than 90. So, we may be more interested in

trying to study, what is the percentage, what is the number of such cruises having more than 130 and less than 90 or between say for example, 90 and 130.

(Refer Slide Time: 27:14)

**Frequency data:
Tabular representation**

India at a glance (% of GDP)	Year			
	1983	1993	2002	2003
Agriculture	36.6	31.0	22.7	22.2
Industry	25.8	26.3	26.6	26.6
Mfg	16.3	16.1	15.6	15.8
Services	37.6	42.8	50.7	51.2
Pvt Consump	71.8	37.4	65.0	64.9
GOI consump	10.6	11.4	12.5	12.8
Import	8.1	10.0	15.6	16.0
Domes save	17.6	22.5	24.2	22.2
Interests paid	0.4	1.3	0.7	18.3

Note: 2003 refers to 2003-2004; data are preliminary. Gross domestic savings figures are taken directly from India's central statistical organization.

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

Frequency data or our tabular representation can be given and as another example. So, say for example, we consider India at a glance in is all economic parameters. So, we have the percentage of GDP which is given along the first column which is the agriculture, industry, manufacturing, services, private consumption, government of India consumption, import, domestic saving interest rates and so on and so forth.

This is for the tribe year 2003 and 2004 and are the based on that we are trying to find finalized. So, in 1983 the agree agricultural consumption is given as 30 c 6.6 percentage. And it goes on to 2003 as 22.2 percentage in the interest paid, it is basically 0.4 for 1983 and it goes to 2003 which is 18.3. So, if you see at a glance, you can find out what is the percentage change which we have.

(Refer Slide Time: 28:10)

Frequency data
Diagrammatic representation

This is the most commonly used for representing time series. The line diagram (also called histogram) is a graph showing the relationship of the given variable with time. There may be three types of line diagram, for the scales used for both the axes of co-ordinates may be arithmetic (or natural) scales, or one of them may be arithmetic and the other logarithmic, or both may be logarithmic. A line diagram where the vertical scale is logarithmic but the horizontal scale is of the ordinary arithmetic type is called a ratio chart or semi-logarithmic chart. When both the vertical as well as the horizontal axes are logarithmic then the chart is called the doubly-logarithmic chart.

Data Analysis & Decision Making R. N. Sengupta, IIM Dept., IIT Kanpur 21

Frequency data or diagrammatic representation can be also be utilized. So, this is the most commonly used for representing time series the line diagram also called the histogram is a graph showing the relationship of the given variable with time. So, they are made three types or line diagrams for the scale used for both the axes and on the coordinate system they can be in arithmetic, on the natural scale. One can be arithmetic and other can be in the logarithmic scale, and the third way of representation can be logarithmic to logarithmic scale.

A line diagram where the vertical scale is logarithmic, but the horizontal scale is of the ordinary type is called the ratio chart or the semi logarithmic scale, when both the vertical as well as the horizontal axes are logarithmic then the chart is called the w logarithmic scale and we will try to study that in more detail later on. So, with this we will I will end the first class and continue discussion about more about data representation and quickly go into the concept of probability and so on and so forth that we are in we will versed and a good position to start off the multivariate statistical critics. Have a nice day.

Thank you very much.