

Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 01
Introduction to Statistics and Data

Good afternoon friends, I am back with a new subject this time, the name of the subject is business statistics, this course is very important, it is being taught in several engineering institutions and in almost all management institutions. Now, this course is specially meant for BBA students, MBA students, there are several students who are pursuing B.Com and they may also be requiring topics which I am going to teach in this particular course.

This course is specially important for part time professionals who are working in different industries, who cannot come to educational institutions and I will be covering lots of techniques which these professionals can use in their daily lives. Now, many of you would be knowing about statistics, this is; this course is not exactly on statistics but it is more of application of statistics into different business scenario.

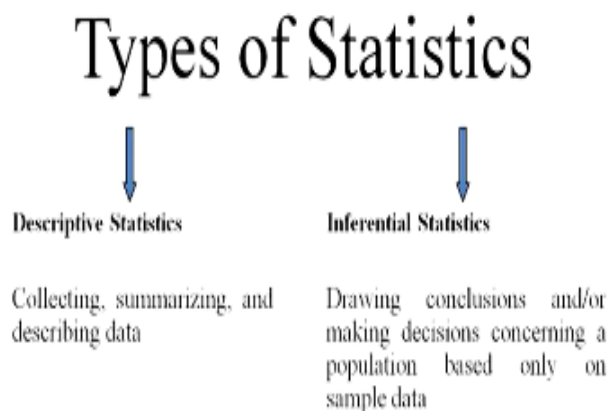
So, statistics is what; at the end of the day, we have got certain numbers and we transform those numbers into some useful information and information should be actionable, now there are lots of information available but many of them are not actionable, so we need to have some data, we need to process that particular data and come up with certain information which are actionable. It is basically methods for processing and analysing numbers.

So, let us say if you have got sales data of a company of last 1 year, so how you are let say, forecasting this sales in next particular month or next particular year. The methods for helping reduce the uncertainty inherent in decision making; now what makes a good manager and how he is different from a bad manager, the most important point is how managers are making decisions. So, if a manager is taking right decision at right time on the basis of data which he has got or she has got.

And after analysing that particular data set how he is coming up with information and how that particular manager is taking action, if you look at different types of statistics, basically you have got 2 broad categories, you have got descriptive statistics and you have got inferential statistics. Now, if you look at descriptive statistics, you are describing something, let us say you are describing a group of students let us say you are saying, I have got 30 students in my class.

So, you are describing that particular class or let us say this product has got height of 5 cm or 5 feet or whatsoever so you are describing certain characteristics of a particular either product or individual or group or organisation.

(Refer Slide Time: 04:13)



So, in descriptive statistics generally, we collect data, we summarise them and we process data, so this is important part of descriptive statistics but if you are coming up with certain inferences on the basis of descriptive statistics then that is called inferential statistics. So, what you generally do in case of inferential statistics, you infer some information from a sample of data and you try to know the characteristics of the population on the basis of sample.

So, basically there are 2 types of inferential statistics and we have got estimation and hypothesis testing, so we will deal with these 2 topics in inferential statistics. Now, why to study this particular subject; business statistics? As I said manager's role is to make decisions, it is not only managers but top bosses always make decisions be it director of any institution, let us say CEO

of any company or let us say any government officer or any army officer, so these people keep making decisions.

And decisions should be based on certain facts, certain data and once you processed those data, you come up with some actionable information, so that is way decision makers use these data and after processing data, they make decisions.

(Refer Slide Time: 06:13)

Why Study Statistics?

Decision Makers Use Statistics To:

- Present and describe business data and information properly
- Draw conclusions about large groups of individuals or items, using information collected from subsets of the individuals or items.
- Make reliable forecasts about a business activity??
- Improve business processes

So, draw conclusions about large groups of individuals or items using information collected from subsets of individuals or items, as I said in case of inferential statistics, you collect data from sample and you try to know about population characteristics, so let us say, you want to know the height of your students who are there in your class, so you are not going to measure height of each and every student.

So, what generally you to do; you take a sample, you take height of that particular selected sample and then infer that the height of the entire class would be this much, let us say 5 feet, 4 inches or whatever it is, so that is inferential statistics. Make reliable forecast about business activities, now this most important aspect as far as business statistics is concern; we want to forecast so many things.

For example, being a company what you would like to forecast, certainly you would like to forecast what would be the demand of your product in next year or next month or next quarter whatever it is, is not it? So, forecast sales you need to forecast how many people you would be requiring next year in your organisation or in your company, so you can forecast your manpower, you can forecast how much raw material would be needed next year.

You may forecast how much money would be required in next year, so forecasting is one of the areas where you can apply statistics on the basis of past data, you can process data and you can come up with some number for next year's forecast and at the end of the day, it improves business processes. Now, there are different business processes, let us say purchasing is one of the processes.

So, purchasing, you get marketing, you get human resource management, let us say you have got after sales service, so once you have got data you can make decision in all these areas, so statistics can be used for several; can be used in several areas, can be used in business memos, business research as I said you can do lots of research, if you have got data and data can be from several sources.

(Refer Slide Time: 09:08)

Why Learn Statistics?

So you are able to make better sense of the ubiquitous use of numbers:

- Business memos
- Business research
- Technical reports
- Technical journals
- Newspaper articles
- Magazine articles

You can prepare technical reports, technical journals, newspaper articles, magazine articles, so all these are different applications of business statistics. As far as this particular course is concerned,

I am going to talk about introduction, data collection and presentation of data, so most important, this is the very first point in this particular you know surveying is the data collection.

So, how you are collecting data, there are different methods, how you are presenting data, there are again hundreds of methods of data representation, what are different measures of location and dispersion, so I will be teaching you in brief about different methods of central tendency, how to calculate central tendency, what are different methods of dispersion measurement and so on.

Probability and probability distributions, numerical descriptive measures and basis probability, discrete and continuous probability distributions, so I will be teaching you what are different probability distributions like binomial distribution, Poisson distribution, normal distribution so on, sampling and sampling distribution, another very important topic in this course, how you are choosing samples, what should be the sample size.

How to collect data from samples, which sampling method is to be used whether it is; it would be probability stick sampling or non-probability stick sampling, then I did discuss about inferential statistics, there are 2 types of this statistics as I have already mentioned; descriptive and inferential.

(Refer Slide Time: 11:00)

- Confidence interval estimation ✓
 - One sample tests and hypothesis testing
 - Two sample tests and ANOVA
 - Chi-Square tests
 - Simple linear regression
 - Multiple regression basics
 - Multiple regression analysis and problems
 - Forecasting analysis
-

So, if you look at confidence interval estimation and one sample and hypothesis testing, so these are different inferential statistics, this would be estimation and hypothesis testing and there are different types of hypothesis testing, you can have 1 sample, 2 sample, 3 sample and so on, then chi-squared test, simple linear regression, multiple regression then we have got forecasting analysis.

So, these are couple of points or let us say topics which I would be covering in this particular course. Let us come to the first topic of this particular course, its introduction and data collection, so before you are going for data collection and analysis, you must know some basics about data, so data; what is data? Data is in fact there are several people have defined data in their own ways and there are certain definitions I would be sharing with you.

(Refer Slide Time: 12:30)

Data: Are collection of any number of related observations

Data set: A collection of data is

Data point: A single observation

Raw data: Information before it arranged and analyzed

Data: Information + Noise

So, data are collection of any number of related observations, just look at this observations, so what could be the observations? I can say that there are 3 people absent today in my organisation that is data, the number of telephone booths installed by the personnel of telephone department that is data, number of buses going from let us say Bombay to Delhi that is data, so all these observations are nothing but data.

When you have got collection of data, it becomes a data set, so let us say today's temperature is let us say 41 degree centigrade, so this is just one data point, if I write temperature of last one

week, then it would become a data set right, so it is a collection of data is data set, right. Data point as I said just single observation, so temperature of Roorkee, today is let us say 41 degree centigrade, so that is just single observation.

Raw data; so information before it is arranged and analysed that is raw data, so let us say if you have got today's temperature and yesterday's temperatures, so 41 degree is temperature yesterday and 41 degree today, so just a data set having 2 data points, right but that is just the raw data because we have not analysed that particular data set having 2 data points, right. So, let us say if I take mean, then it becomes analysed data, right.

One of the ways of defining data is information plus noise, now every data set will have some information and some noise, so being a manager or being a decision maker, your job is to extract information from available dataset, so either extract information or you just remove noise is one and the same thing and it is very difficult let me tell you, so data is information plus noise, so you should remove noise from data.

Because data generally speak and you should hear what ever data say, a good manager always watch the data, listens to data and then take appropriate decision.

(Refer Slide Time: 15:40)

Example of raw data:

High school and college CGPA	
HS	College
3.6	2.5
2.6	2.7
2.7	2.2
3.7	3.2
4.0	3.8

Let me give you an example of raw data, so here is; you have got 4 data points, so high school and college CGPA of 5 high schools and 5 colleges, so let's say CGPA of first high school is 3.6 and college CGPA is 2.5 similarly, here is CGPA is 4, here it is 3.8, so can you come up with certain inference out of these 2; 2 sets of data; high school data and college data? Because there is no co-relation, right.

Because let us say when CGPA is 3.6, here in high school it is 2.5, when it is 4, it is 3.8, right, so this is here it is decreasing but here it is increased, is not it? It was 2.5, now it has become 2.7, so again it has increased right, now here it is increased from 2.6 to 2.7 but it has decreased, so you do not know how these two data sets are correlated, right however, there would be certain correlation definitely but this is just a raw data, okay.

(Refer Slide Time: 17:21)

Criteria/Tests for Evaluating Data

Criteria	Issues	Remarks
Specifications & Methodology	Data collection method, response rate, quality & analysis of data, sampling technique & size, questionnaire design, fieldwork.	Data should be reliable, valid, & generalizable to the problem.
Error & Accuracy	Examine errors in approach, research design, sampling, data collection & analysis, & reporting.	Assess accuracy by comparing data from different sources.
Currency	Time lag between collection & publication, frequency of updates.	Census data are updated by syndicated firms.
Objective	Why were the data collected?	The objective determines the relevance of data.
Nature	Definition of key variables, units of measurement, categories used, relationships examined.	Reconfigure the data to increase their usefulness.
Dependability	Expertise, credibility, reputation, and trustworthiness of the source.	Data should be obtained from an original source.

Now, there are different sources of data, you can classify them as primary source and secondary source, so whenever you have got data whether you have collected it from primary source or from secondary source, you should evaluate data, you should test data whether that particular data set is appropriate or not and there are different criteria on which you should evaluate data. First one is specification and methodology; let us say if you have taken secondary data.

Secondary data is data which someone else has collected for some other purpose and you are just referring for your use, it might be useful for you, it might not be useful for you, so if you are

looking at let us say secondary data, then you should look at specification methodology of that particular secondary data. So, how what was the method of data collection, when you are referring to secondary data, what was the response rate?

Let us say somebody else collected data on let us say 100 people and the response rate was let us say 10%, so is that data set useful for your study, so you to check it, right, so what was the response rate, how the analysis was done on secondary data sampling, what was the sampling methods used, what was the sample size, how many questions were there, whether the questionnaire was pilot tested or not?

Who did field work, who controlled fieldworkers and so on, so this is very important point, specification methodology as far as evolution of data is concern, so at the end of the day what you want; you want reliable data, right, if it is a secondary source, even if it is primary then also you need to look at reliability of data, validity of data and generalizability, so there are 3 important points; reliability, validity and generalizability, right.

Errors and accuracy in secondary data, so you should look at the errors which might be there in secondary data, what was the research design used for collecting secondary data, whether it was exploratory research design or conclusive research design, how data were collected, whether the data were collected let us say in person or over phone or over Internet or in mall intercept or you collected data at home or at office, is not it?

So, look at all those aspects as far as data collection is concern, right and you are evaluating secondary sources of data right, so assess accuracy by comparing data from different sources, let us say suppose if I ask you what would be India's GDP next year in 1890, so I should get this particular information from several sources, is not it; rather than just looking at one particular source.

So, I can look sources like let us say centre for monitoring Indian economy database, I can look this information on RBI's website, I can look at this information on the Ministry of Finance website, I can look at this data on simple just by Googling, I can look at these data by talking to

let us say Finance Minister, is not it, so there are several sources and I should have data from different sources and then try to come up with a particular solution or a particular number.

Then the next one is currency, how updated that data is; is the time lag between the data collected and data published, how frequently the second resources are being updated and we should see if you look at the census data in India, the census data are updated every 10 years, right, so how the forms which are supplying secondary data, how that firm is updating secondary sources data.

Then look at the objectives; what were the objectives of collecting secondary data? Because you may or you may not use secondary data but you may get some information from secondary data for your own research, so the objective will tell you whether the data which you are collect which you are taking for your research are relevant, having any relevance or not with your research.

Then look at the important key variables, we will see what is the meaning of variable, unit of measures, measurement, categories relationship examined and so on, so reconfigured the data to increase their usefulness, so this is the objective here. Look at dependability; when you are looking at secondary source, you should look at the credibility of that particular source, the reputation of the organisation from where you are taking secondary data, how trustworthiness in that particular sources.

And it should be as far as possible data should be obtained from an original source, right so you may have different data and data again as I can be let us say numeric data, nonnumeric data or let us say structured data or unstructured data, you may have image data, you may have text data, you may have video data and so on, so there are different types of data, so you should look at these issues in secondary sources, okay.

(Refer Slide Time: 23:47)

Double counting example: Truck association “75 % of the items you receive come from trucks”. Does it mean 25% Rail, Airplane, Ship.

Now, there is something called double counting example, whenever you are collecting data then while collecting data and while analysing data, you should take care of something called double counting. Let us say if the truck association of a country or of a state says that 75% of the items you receive come from trucks, just see this, so this is the claim of the truck association that 75% of the items you receive at home come from trucks.

Does it mean the remaining 25% of the time, the use of rail, aeroplane and ship has been there, it would be a wrong interpretation, it is possible that the items which you are getting at your home might have come to your city either by ship or by rail or by aeroplane but the distribution in the city was by trucks, so this is a case of a double counting, right, so avoid such type of inference from data.

(Refer Slide Time: 25:45)

Elements, Variables, and Observations

- The elements are the entities on which data are collected.
- A variable is a characteristic of interest for the elements.
- The set of measurements collected for a particular element is called an observation.
- The total number of data values in a data set is the number of elements multiplied by the number of variables.

Now, let us look at couple of other points related to data, so I will be defining elements, variables and observations, so elements are the entities on which data are collected, let us say I want to collect data on height of students, so those students are nothing but elements, right, so elements are the entities, so entities can be a group of people, a group of organisation and individual person, a product or a group of product, so all these are elements.

Variable is a characteristics of interest for the elements, so let us come to the entity as a group of students, so let us say I am interested in their height, so this height would be nothing but a variable okay, let us say height of a student or weight of student or let us say marks of the student all these are nothing but variable, so you are interested in certain characteristics of elements right.

The set of measurements collected for a particular element is called observation, so let us say if I measure height of a group of students, let us say 5.5 inches, 6.1 inches and so on, so that would be called observations, right. The total number of data values in each in a dataset is the number of elements multiplied by the number of variables is very simple.

(Refer Slide Time: 26:59)

Data, Data Sets, Elements, Variables, and Observations			
Observation		Variables	
Element Names	Company	Stock Exchange	Earn/Share(\$)
Dataram	AMEX	73.10	0.86
EnergySouth	OTC	74.00	1.67
Keystone	NYSE	365.70	0.86
LandCare	NYSE	111.40	0.33
Psychomedics	AMEX	17.60	0.13

So, let us take an example of data, data set elements, variables and observations, so here there are different companies let us say, Dataram, LandCare and so, so all these are nothing but different companies and we will call them elements, right as I said group of students, group of products all those are elements, right and the variables are I am interested in let us say earning per share of these particular company is 0.86, so this is a variable, annual sales is this much, this is another variable.

So, I have looked at earning per share from stock exchange of this particular company, so AMEX stock exchange, so there you may have different stock exchanges, right New York Stock Exchange, is not it, so these are different observations and this is nothing but dataset, so and this is data point, every single observation is data point. Let us look at once again what we have done in today's session.

We have talked about this; the course and for whom this course is has been designed, what are the contents of this particular course, we have seen different types of statistics, we have seen descriptive statistics and differential statistics, we have also seen different definitions of data, data points, dataset, observations, elements, entities and so on. So, with this let me complete today's session.

In next session, we will carry forward this particular topic which is introduction and data collection, thank you very much.