

Lecture 06
Introduction to Descriptive
Statistics - Objectives, Steps

Welcome to the, lecture on the course descriptive statistics with R software. Now you may kindly recall, that in all the earlier lectures we had, discussed and we had an idea, that how our software is going to help us in, different types of computations. Now from this, lecture we are going to start the discussion, on the topics of Statistics. But, here I would like to tell you one thing or I would like to clarify one thing, the topics in descriptive statistics whatever I am going to consider here, they are pretty elementary and I will try to, go to the depth, as much as possible, under the given time frame, but my idea is not really here to teach you statistics, my idea here is that, most of the topics you will see, you know? And my objective is that, I would like to make you comfortable, that in case if you want to use the our software for the computation of those topics, then you should be comfortable, you should be confident, once you are confident in handling the basic, topics I am, sure and I am confident, that there should not be any problem in handling the, the advanced topics in statistics. Besides this thing, people are using these statistical tools, very often. But sometime, they don't know why they are using it; sometimes they don't know what is the interpretation of different quantities. So, that will be my another, objective on which I would try to discuss here, that whatever the, the tools of descriptive statistics I am, going to handle, I will try to discuss about their concept, their implications and their competition using the our software. Right? So, in this, lecture I am definitely not going to use any our software for the competition but, I would try to give you, an overview, that what is descriptive statistics and how it, helps and what are the ways, in which it can give us different types of hidden information and a given set of data.

Refer Slide Time :(2: 49)

Objectives of Statistical Analysis

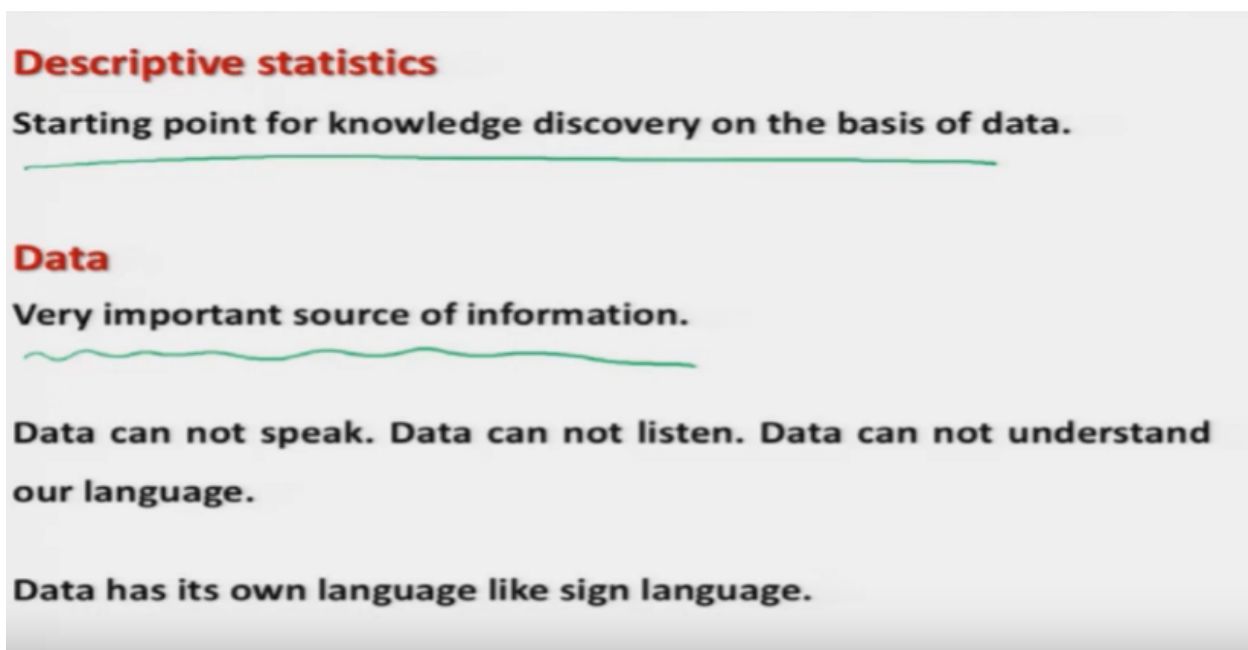
What is Statistics?

It is

- **Science of turning data in information for decision making.**
- **Collection, analysis and drawing inferences from numerical facts, referred as data.**

So, only I will be considering on the basic, aspects in this lecture and possibly in the next lecture also. So, one of the basic fundamental question comes here, what is really statistics? So, I would say simply, that statistics is a science, which turns the data, or which tries to take out the information contained inside the data and converts, it into a form which is useful for making a decision, this decision can be at, your office level, at policy formulation, for forecasting, at country level or say anything else. So, what are we going to do here? We are trying to collect the data, we are trying to analyze the data and based, on that we will try to, make some statistical inferences and those, inferences are drawn, from the numerical facts which we are going to call as, as data. So, data is a very, important thing in statistics and it gives you, some information and this information is hidden, inside it and what is descriptive statistics?

Refer Slide Time :(4: 03)



Descriptive statistics
Starting point for knowledge discovery on the basis of data.

Data
Very important source of information.

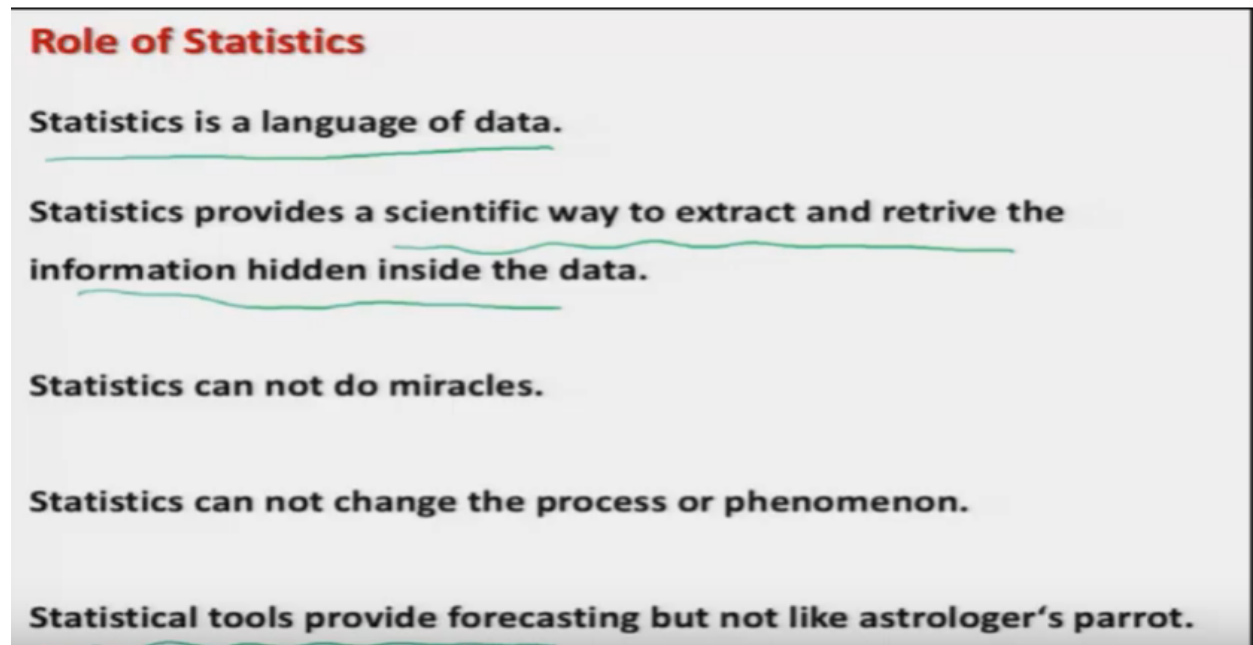
Data can not speak. Data can not listen. Data can not understand our language.

Data has its own language like sign language.

This is, essentially the starting point for knowledge discovery on the basis of data. Yes, the discovery of knowledge can be done from different types of ways, by looking at the picture, by reading some subject and similarly aesthetics, also gives us a tool to, discover the knowledge that is contained inside the data. So, data is essential a very, important source of information, data contains many information inside 8but, the problem is the following, if I know something, then I can speak and I can inform you, but data cannot speak, data cannot listen, data cannot understand the language what we speak, for example if I am, speaking Hindi or English language possibly you can understand it, but if I try to speak in a language, which you don't understand, then I will not be able to transfer the knowledge. So, similarly data has its, own language for example, you have seen a first somebody, has some, problem in speaking or hearing then there is a, Chinese language and that language can be, understood only by those, people who understand it. So, similarly data also, has a, language and which is based on, different types of symbols,

notations and interpretation. So, our objective here is that, that we want to know the tools, by which we can draw, the information contained inside the data using the tools of descriptive statistics.

Refer Slide Time :(5: 53)



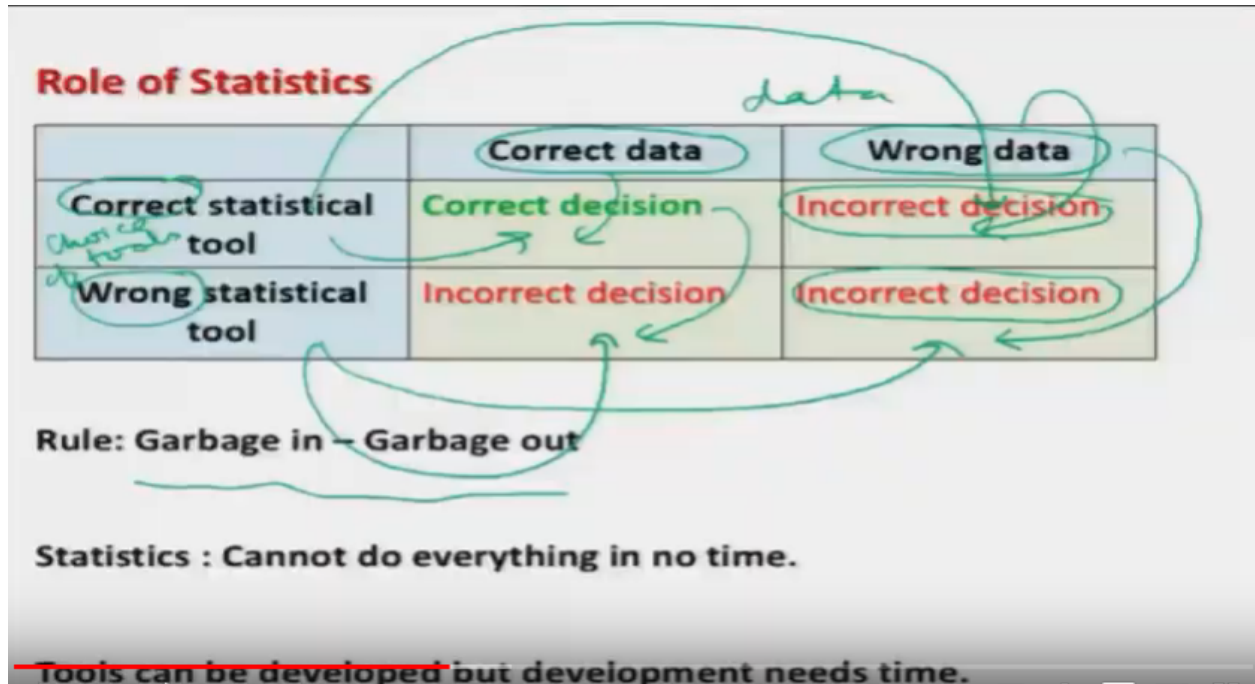
Role of Statistics

- Statistics is a language of data.**
- Statistics provides a scientific way to extract and retrieve the information hidden inside the data.**
- Statistics can not do miracles.**
- Statistics can not change the process or phenomenon.**
- Statistical tools provide forecasting but not like astrologer's parrot.**

So, essentially I can say here, that statistics is a language of data and it provides a, scientific way to extract and retrieve the information hidden, inside the data. And remember one thing; its taxes cannot do any miracle, sometime if you try to see, you might have heard, that people try to make different types of joke, different types or comments, on the statistic that is that 6 lies or something like this. So, I would like to inform you that, it started 60 never lies. Right? The statuses is simply based on the data and now it is our, capability that how, much we can retrieve the information in a from inside the data, for example, if somebody is speaking to us, in a sign language then it depends, on our capability that how, much I can correctly interpret. It so remember one thing, that aesthetics also cannot, change the process or the phenomena, whatever process is happening that will happen and as a statistician, I am not allowed to, alter or change the process, I simply have to collect the data, on the basis of the process which is happening, which is continuing and on that basis I have to take a call, I have to take a decision, that which of the tool is most appropriate tool in this situation, to draw or to extract the information hidden inside the data. Right? Ok and definitely, the inferences what we are going to draw on the basis of a statistical tool, they are going to use for different purposes, for example one of the basic purpose is forecasting. So, for example stat scale tool provide, forecasting but remember one thing, this is not like a astrologers parrot, that the parrot chooses a, chit and then one reads it that what is my future, it does, forecasting but on the

basis of some scientific principle and the principles of statistics. Now whenever you are, coming to this aspect, that you have to choose, a tool, on a data set, to extract the correct information, then in this process I can, divide the entire canario, of decision-making into, two parts.

Refer Slide Time :(8: 19)



One part I would say here, suppose my data is here and there are two option, that the data is correct and the second option is, this data may be wrong, what do you really mean by data is correct and data is wrong? Suppose I want to know, the average height of the students in a class, then obviously, I have to collect the data on the height. But suppose I am, trying to collect the data on the weight and based on that I am, trying to infer the data on the height, then it is not appropriate. Right? So, in that sense I am, trying to say, that the data has been chosen correctly, which is matching, with the objective of the study and second aspect is, choice of tool, there are many, many state scale tool, which are, available and we are going to study them in the future lectures, in the further lectures. But the main thing is this, one has to choose, what is the correct tool to diagnose, the problem and to solve the problem, just like, if you go to the shop of a medicine or if you go to a doctor, there are thousands of medicine. But, what doctor does? Doctor try to; decide that which, medicine is most suitable for the given problem. So, similarly in statistics also, we have many, many tools and we need to make a decision that, which tool is going to be appropriate, to draw the correct aesthetical inference, for a given problem. So, I know, how the choice obstetrical tool, can be correct or the choice of statically tool can also be, wrong. Now there are four

options, first option is this, suppose I am, trying to choose the wrong, stat scale tool, over a wrong, set of data. Then in, this case the decision is not going to be the correct, The next option is this, means I can choose the wrong, start school tool and the data is correct, even in this case, I'm trying to use the wrong tool over a, correct data I will get a, incorrect decision and similarly, in case if I try to take here the, wrong data and if I try to take here the correct aesthetical tool. So, using the correct start scale tool, all are wrong data will also, give us the incorrect decision. Now, the last option which is correct is that, I have to use, the correct stat scale tool, over our correct data. So, this is the only option, though so, unless and until you try to choose the correct esthetical tool and you collect your data correctly, you will not be able to get the correct aesthetical inference out of it. The rule is very simple, garbage in, garbage out. Right? And, one thing we wish I would like to inform is that, sometime if people do come to us and they ask us to, do the Statistics very quickly, for example I have to submit my thesis tomorrow, I have this data please try to help me, write, well, it is not, so simple at that stage. Because you are the one who has understood the entire process and as a statistician, I have not, been told about your process. So first you need to explain me, the entire process then I will try to understand the data generating process and only then, I can do something and it is also possible, that the type of tool, which you need, that may not always be available, but it needs to be developed. So, statistics always, need some time to understand the phenomena. Now, another popular question with which I, am asked that when I come to the aspect of descriptive statistic there are two types of tools. One is graphical tools and C and another is analytical tool.

Refer Slide Time :(12: 23)

Descriptive statistics

Several components and tools- Graphical, analytical

Graphical tools- various type of plots

- 2D and 3D plots,
- scatter diagram
- Pie diagram
- Histogram
- Bar plot
- Stem and leaf plot
- Box plot etc.

Use appropriate graphics.

So, there are different types of graphical tool, like as, to dimension, three dimension plots pie scatter diagram, pie diagram, Histogram bar plot, stem leaf plot, box plot and etc there is a long list. In this case people do ask me, which of the graph is more suitable or they also, feel that if they try to use more, number of graphs, then their analysis is going to be better, I would say this is only a myth, you simply have to choose the correct, graph and you have to, use the correct number of graphics. So, the appropriate choice of graph and appropriate number of graphics, that is only going to help you in getting the correct information and similarly when I come to the aspect of say they say analytical tool,

Refer Slide Time :(13: 13)

Central tendency of data	Dispersion in data
• Mean	• Variance
• Median	• Standard deviation
• Mode	• Standard error
• Geometric mean	• Mean deviation,
• Harmonic mean	• Absolute deviation
• Quantiles etc.	• Range etc.

Their analytical tool, I can there are different aspects, on which, we try to analyze the data, for example I would try to, find out what is the central tendency of the data? What is the variation in the data? And what is the structure of the data? And what type of relationship are existing inside the data? So, for example, if when I come on the aspect of measure of, central tendency then we have different tools mean, median, mode, geometric mean, harmonic mean, quantiles, etc and similarly when we are trying to, understand the variability in the data, then we have different types of tools, variance, standard deviation, standard error mean deviation, absolute deviation, range, etc. So, you can see here, in this case, there are two aspects, one is the central tendency of the data and another is desperate or the variation in the data. Now, in case if you want to study the central tendency of the data, then and then out of this list, you have to choose the appropriate tool and similarly in case if you want to study that, nature of the variation in the data, then you have to choose the appropriate tool. And similarly in case if you want to find out, what is the structure of the data in terms of, symmetry peakedness etc,

Refer Slide Time :(14: 26)

Analytical tools

Structure of data	Relationships in data
<ul style="list-style-type: none">• Symmetricity ✓• Skewness ✓• Kurtosis etc. ✓	<ul style="list-style-type: none">• Correlation coefficient ✓• Rank correlation ✓• Multiple correlation coefficient ✓• Partial correlation coefficient ✓• Corralation ratio ✓• Intraclass correlation ✓• Linear Regression ✓• Non linear regression etc. ✓

Then you have to, choose a proper tool for symmetry and then there are concept of skewness, concepts of kurtosis, these concepts are going to give us more information, that what is the structure of the data? And then, another aspect can be, if I have, a data on say more than two aspects something like height or weight or say height, weight and age, then there may exist some relationship in the data also or there may exist some, coherent structure inside the data. Then in order to study those aspect we have the tools of, correlation coefficient, rank correlation, multiple correlation, partial correlation, coefficients correlation ratio, intra class correlation coefficient, linear regression, nonlinear regression etc. So, there is a long list.

Refer Slide Time :(15: 12)

Statistical thinking and Methods

Which of the tools to be used – Graphical or analytical?

Use both types of tools.

Graphical tools provide a visualization – First hand information.

Analytical tools – Quantitative information.

Both approaches work together and are inseparable.

So, when we talk about the descriptive statistics, descriptive in statistics is not a tool. But this is the collection of the appropriate number of tools, that may include the graphical tools, as well as, that may include the analytical tool and the choice of analytical tool also, depends what exactly do you want to study? Many times people, do come to us and they ask us, sir, can you please do some static analysis on my data, in that case I would always, request them please, let me know what really you, want to know from this set of data? And based on that, I am going to take a appropriated decision and I'm, going to decide that which of the statistical tool, can give an answer to your query and then I would try to use it, so another question crops up here, that which of the tool is a better option, graphical tool or say analytical tool, my suggestion is that, please use both of them, because if you try to see, this descriptive is Statistics, is the point of starting of a name, for this analysis, what you have in your hand? You simply have a set of data, data are some numerical values. So, you can always imagine that in front of you, there are 20 values, there are hundred values, there are 2,000 values or they can be two million or so two billion values. And all those values are, sitting silently and you are the one, who is going to, start the knowledge discovery on the basis of given set of data. Right? So, I would say, don't make a rule, but depending on the condition, try to use both types of tool, later on in these lectures, I will show you that, how the graphical tools and, and how these analytical tools can be used, under what type of condition, how they can be computed on the basis of R software. Now, what is the difference between, between the use of graphical tool and an analytical tool? Graphical tools provide us a, visualization. This will give us a, first-hand information and what about analytical tools? They will give us the, information in quantitative form and they will give us the quantitative information. So, graphics, will give us information, but we have to look into this and then we have to draw a proper esthetical inference. And this analytical tool, will give us a number, which we have to interpret to make a correct hysterical inference. So, I would say, usually, are in most of the cases, this graphical tools and analytical tools, both work together because, the process is the same, data is the

same and data is propelling you, that please use, only the graphical tool or please use, only the analytical tool, this is only you, who is going to take a call? That whether graphical tools have to be used or say, these analytical tools have to be used. So, please try to make a, appropriate decision keeping in mind the objective of your study and the type of information which is, contained inside the data. And in statistics, why do statuses comes into picture? It's Texas comes into picture.

Refer Slide Time :(18: 40)

Statistical thinking and Methods

Both – graphical and analytical tools – work together in a system of interconnected processes.

Variation exists in all processes.

Understanding the extent of variation and reducing it are the keys to success.

Because variation always exists in all the process. What do you mean by variation? For example, if I say, suppose you try to take a plot and you try to put, say this hundred grams of seeds and say after a month, you will get a crop and suppose you will get, one kilogram of seeds. Now in case, if you try to repeat the same thing, try to use the same plot or the same sizes or plot and put the same quantity of seeds, do you think that, in all the plots, you are always, going to get exactly, one kg of field, this is practically very difficult, there will be some difference, one plot may be given you one kg, another plot may be given you one point, one kg and say another, plot might be giving you 900 grams and so on. So, the variation always, exists in all the process and in statistics our, basic objective is that, that we want to, understand the process of this variation, we want to control this variation. And to draw a statistical inference out of the data, with minimum variability. So, this is, one of the basic objective, so in statistics or in descriptive statistics what we really want to do, we have a setoff data, now we are going to use that data, on a aesthetical tool, that may consist of say analytical tool, as well as graphical tool. Now I will be getting

some information from the graphical tool, I will be getting some information from the analytical tool and now, this is my responsibility, that I have to combine, the information coming from, both the aspects together and I have to convert it, into our, piece of information, which is useful, which is interpretable or that can be conveyed, to the experimenter, who might not have any knowledge about the statistics. Right?

Refer Slide Time :(20: 50)

Statistical thinking and Methods

Using the information gained by the tools of descriptive statistics and combining them together to reach to a meaningful conclusion to depict the information hidden inside the data is the objective of any statistical analysis.

Proper interpretation of inferences drawn is important.

Inferences are drawn from the data.

~~Data generating process. Non deterministic, Random~~

So, I can say, that using the information gained by the tools, of descriptive statistics and combining them together, to reach to a meaningful conclusion, to depict the information hidden inside the data, is the objective of any stats color analysis and proper interpretation of, those outcome is very, important and all these outcomes, all these inferences are made only, on the basis of data. So, P next question come, how this data is coming? So, there are two processes, one deterministic process and say another is non-deterministic process, deterministic process means, you know? The outcome in advance. But non-deterministic process are, where you really do not know, the outcome in advance. So, in statistics whenever we are, calling or when we are and whenever we are trying to understand the data generating process, the data generating process is always random or say, non-deterministic and that is why, the role of statistics comes into picture, once there is no, random variation things will become purely, mathematical.

Refer Slide Time :(22: 06)

Why collect data?

To verify theoretical findings.

Draw inferences just on the basis of collected data.

Developing statistical models, which can be further used for policy decisions, classification, forecasting etc.

So, a simple question that arises here, why should we collect the data? So data is collected with different types of objective. First is, to verify the theory all findings, for example suppose if I say, in children the height increases as the, weight increases or the weight of that child increases as the height increases, suppose that is my theoretical finding and ever and if I want to, verify it whether, this is, really happening in real life or not? So, I need to collect the data and I need to verify this finding, secondly I have some objective and I really want to know, the outcome of that process, so I have to, collect the data, which is being generated from that process and then I have to use a statistical tool, to the correct aesthetical outcome and remember one thing, the inference what we are going to draw? That is, just on the, basis of the collected data, you cannot argue, that some statistical inferences, is coming and which is, from some other source, beyond the data. So, particularly when we are talking of the tools of descriptive statistics, we try to speak of information or we try to report the information, which is coming, only from the given set of data. And yes, the information which is coming from this data, that we try to convert it, in the form of statistical inferences, which is further use, in the development of state scale models, which are used for policy, decisions, gradually a classification, forecasting and many, many other things possible.

Refer Slide Time :(23: 55)

Steps Involved

- Identify objective(s) of Statistical Analysis
- How to get data?
 - Laboratory experiment, ✓
 - survey, ✓
 - primary data, ✓
 - secondary data ✓
- Use appropriate statistical tool
- Correct, valid and meaningful interpretation of the result.

Now, in case if you want to, make a study on a statistical experiment, then what are the steps involved. Right? The first objective step, that please identify the objective of the astatically analysis, which is missing in, most of the cases in my experience. Right? People simply try to collect the data and after collection of the data, many times they try to, decide what type of statistical inferences they can draw from this data? Well that is not bad, but at least my suggestion is that, before you collect the data, please try to, decide the objective of your study and try to ask, why I am collecting the data and based on that you have to take, further steps. Okay? How to get the data, the data can be obtained from a laboratory experiment, from a survey, from some primary sources, or from some secondary sources, called as, 'Primary Data or say, Secondary Data'. But whatever is the data, I am not bothered about that, my objective is this, I have to use the, correct statistical tool and I believe, that by using the correct aesthetical tool, I can get a correct, correct, valid and meaningful interpretation of the result,

Refer Slide Time :(25: 16)

Observations

The units on which we measure the data are called observations.

For example, number of persons, cars, monthly expenditure on food etc.

that is my belief and with this objective I am, moving forward. So, the next question comes, what is an observation? So, the unit, on which we try to measure the data, that is called an, 'Observation'. What did this mean? Suppose I want to measure the height, so first I have to decide height of children or say height of elders, suppose I decide, that I need to measure the height of the children between the ages, say 5 years, to 7 years. So, what I would say? That I will try to collect is, some children whose ages are between five years and seven years and then I will try to, record the heights of those children. So, the heights of those children, which will be some numerical value and they will be called as an, 'Observation'. So, similarly in case if I want to find out the number of persons, number of cars, monthly expenditure, on say food in any family, then these are also my observations, which are trying to cater, to some objective of statistical analysis, next definition which I would, like to discuss here, what is called a population? The collection of all, the unit is called a, 'Population'. For example, in the earlier, example when I wanted to, the data, on the height of children between five and seven years, you are trying to collect the data, only on some of the children. But do you think, they are the only children, no there are many, many children in that city, in that locality, in that country. What, what are you trying to do? You are simply trying to choose, some of the children and then you are trying to record the data. So, the collection of all, the data, that can be locality, that can be City, that can be country, that depends on the objective, that is called as a, 'Population of children' whose ages are between five years and seven years.

Refer Slide Time :(27: 30)

Population

Collection of all the units is called population.

Example:

Objective: To find average age of all the female students in class 10 in a school on the basis of a sample.

Population: All the female students in class 10 in the school.

Similarly, suppose I want to find out the average age of the, of all the female students, in class ten, in a school, on the basis of a sample. Then by, population is going to consist of all, the female students in class ten, in that school, that will be my population. But if I want to study, the average age of all the female students in that city, then all the female students in that city, who are studying, in class ten, that will consist of my population.

Refer Slide Time :(28: 07)

Population

Example:

Objective: To find how many female employees have salaries more than the respective male employees in a company on the basis of a sample.

Population: All the male and female employees in the company.

And similarly in say, in another example if my objective is that I want to know, that how many female employees have salaries more than the, male employees in a given company, on the basis of a sample. So, in that case, my population will consist of all, the female employees in the company. From there I will try to choose, a small sample.

Refer Slide Time :(28: 30)

Sample

Sample is a subset of the population.

Selection of observations in the sample from the population is made in such a way that the sample is representative.

Sample is representative.

Representative sample

All characteristics present in the population are also present in the sample.

Now, the next question is what is a sample? So, sample is only a subset of the population, a basic question comes; why, we use this sample? Well, that is the main objective and main advantage of using statistics, we are always interested in finding out a statistical conclusion, which is for the entire population, maybe of country, maybe of city or maybe of village and there will be, large number of people, who have to work, to collect the data, that is very difficult. So, the advantage of statistics is that, that the statistics says, that instead of collecting the data on the entire, population if one can collect the data, on a small fraction, that we call a sample, then on the basis of sample of the data, the Statistics can help, in getting a reliable, esthetical inferences which are going to be valid for the entire population. So, that is why? Collection of sample is very, important in statistics. So, whenever we are trying to collect the data, in a sample, we believe, that whatever are the characteristics, whatever are the features, which are present in the population. They are also, present in the sample for example, you have seen that if you go to a market and you want to buy, some wheat and there is a bag of hundred kilogram of wheat, usually, you will not open the entire bag, but you will try to take a small sample, maybe consist of say this 20 grains, 40 grains hundred grains, you simply try to look at those grains and based on the quality of the grains, you try to make an inference, for the entire bag, which is of hundred kg. So, now, this is my sample, sample is possibly consisting of the grains of wheat, maybe 20 gram, 30 gram, 40 gram and based on that whatever we are going to continue, that is going to valid for the, for the entire bag. It's not even the entire bag, but I will say, the entire wheat, available in that shop. So, that is why? The collection of data, in the sample is very, important and we believe, that the data has been collected in such a way, such that the sample representative. So, this will be our basic, assumption and that goes without saying, that in all his tactical analysis the sample means, sample is representative, what is the representative sample? That means all, the characteristics which are present in the population, are also present in the sample, for example incase if the, quality of the wheat is not so good, suppose there are ten grains, which are infected by some

insects. Then we assume, that in the population also, means a similar type of proportion will continue, in case of the ten percents seats in my sample, in my hand, are not good so we, so I believe, there 10% of the wheat in the, entire bag is also not good. So, this is how we go,

Refer Slide Time :(32: 05)

Sample

Various sampling schemes like

- simple random sampling, ✓
- stratified sampling, ✓
- cluster sampling, ✓
- systematic sampling, ✓
- multiphase sampling, ✓
- multistage sampling etc. ✓

~~are used to obtain a good sample.~~

So, about basic, foundation like that in case if my sample is good, sample is representative, then my statistical inferences are also, going to be good. And there are various ways, in statistics which help us, in choosing the representative sample or, or say character sample so, so we have different types of sampling scheme like simple random sampling, stratified sampling, cluster sampling, systematic sampling, multi phase sampling, multistage sampling etc. And which helps us and guide us, that how to choose the correct, good and representative samples, but definitely this is not the objective of the course and I'm, not going to discuss that what are the different sampling? Procedures, to collect a good data. So, now in this lecture, I have given you a brief background that may not, look very mathematical but, believe me this is, very important for us to understand that what are we going to do, under what type of condition, only then, I will be able to take the correct decision and I have, recorded or I have given this lecture, with this objective only. So, I will try to continue with some more basic definition in the next lecture. And you try to understand this lecture and try to create a foundation inside your mind, to understand the tools of descriptive statistics and I will see you in the, next lecture. Till then, Goodbye.