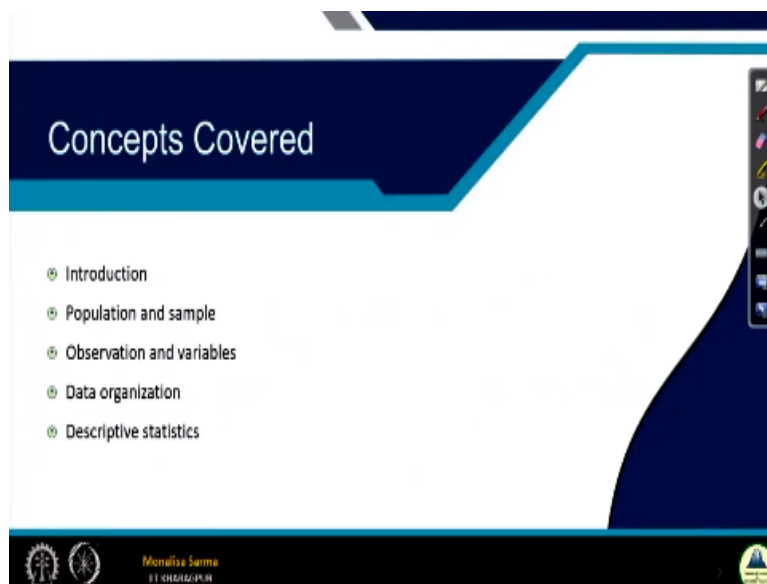


Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology - Kharagpur

Lecture – 02
Introduction to Statistical Methods in Reliability

Hello everyone. So, this is the second lecture on the course statistical learning for reliability analysis. And the last class we got introduced to reliability engineering and this is the introductory class on statistical methods.

(Refer Slide Time: 00:39)



So, in this class basically we will try to learn about statistical methods and we will learn various terms related to statistical method that is like population sample, what is population, what is sample what is observation, what is variable and then we will learn also some data organization tool and then we will go to discuss some descriptive statistics.

(Refer Slide Time: 01:04)

Introduction

When actually people started thinking about statistical methods in reliability?

- Does Japanese "industrial miracle" has role in it?
- Much of the success of the Japanese has been attributed to the use of statistical methods and statistical thinking among management personnel.

Statistical Learning for Reliability Analysis

Monalisa Sarma
IIT KHARAGPUR

So, statistics when we talk it is not a new concept it was there from long back like, but long back it was there, but then it was never used in industry long back like when actually people started thinking of statistical methods in reliability and from the reliability concept and asking. So, does Japanese industrial miracle has a role in it? So, what is Japanese industrial miracle? If you just Google it will be able to say at as in the middle of the 20th century, Japanese were very much successful, where we India or many other countries has failed miserably, where we have failed it is like developing such an environment where we can develop some good quality and reliable product. So, it was in the middle of 20th century Japanese were very much successful very developing very quality product, very reliable product but other countries that involvement was not there. So, as you can develop a reliable and good performance product.

So, much of the success of the Japanese has been attributed to the use of statistical methods and statistical thinking among management personnel. So, now, when we talk about use of statistical methods, like this data has when we talk about statistics definitely we talk of data like so, this using data some collecting data, summarizing data and then reporting data all this was quite prevalent in many other industries from long back, but then collection of data and statistical treatment of this data, these are 2 different issues.

So, before this Japanese industrial miracle people were just collecting data they were collecting, they were summarizing the data and some other format, they were reporting the data and storing

the data for the purpose of result but they were not doing any statistical thinking on those data. So, it has started this Japanese successive, Japanese company this has made people think and then people started visually started using these statistical methods in their industries.

(Refer Slide Time: 03:13)




Now, when we talk of statistics as I told you, the first thing that comes to mind is what does it comes the comes to mind it is like a big, big tables of many numbers and volumes and volumes of figures, figures maybe pertaining to birth, death, taxes. So, that is why because statistics is very much used synonymously with data, this is actually in fact not very wrong also, because when we talk of statistics, Statistics basically it largely deals with principles and procedures of collecting data, collecting and summarizing and what to say taking information out of this data. So basically, when we talk of statistical method, so a good idea of what is data is very important. So, first we will go to that what is the data?

(Refer Slide Time: 04:12)


What are Data?

Data: A set of data is a collection of **observed values** representing one or more characteristics of some objects or units.



Example

- > Name of students enrolled in this course
- > Age of students enrolled in this course
- > Institute name of students enrolled in this course
- > Marks obtained by students, etc.




Monalisa Sarma
ET CHHARAPUR

So, data is how we define data, a set of data is a collection of observed values representing 1 or more characteristics of some units, objects or units. Now, like let us take an example of a student when we if the student is a unit in particular, one student is a unit so the different data pertaining to that unit is the name of the students maybe the age of the students, institute name marks obtained by the students. These are the different data pertaining to the unit students. So, all this together, we can tell it is a set of data.

(Refer Slide Time: 04:45)

Example of a Typical Data Set

| Respondent | Age | Sex | Happy | TV-Hours |
|------------|-----|-----|-------|----------|
| 1 | 41 | 1 | 2 | 0 |
| 2 | 25 | 2 | 1 | 0 |
| 3 | 43 | 1 | 2 | 4 |
| 4 | 38 | 2 | 2 | 2 |
| 5 | 53 | 2 | 2 | 2 |
| 6 | 43 | 2 | 3 | 5 |
| 7 | 56 | 1 | 2 | 2 |
| 8 | 53 | 2 | 2 | 2 |
| 9 | 31 | 1 | 2 | 0 |
| 10 | 69 | 1 | 1 | 3 |
| 11 | 53 | 1 | 3 | 0 |
| 12 | 47 | 1 | 2 | 2 |



Monalisa Sarma
ET CHHARAPUR

So now, it is a typical data set. So, it has not surveyed done by for some robot to say for some research purpose when people were interviewed. Today is mainly social science survey where many people were interviewed around 2000 people were interviewed, and they were asked

different questions there were around 70 questions, there is based on their political beliefs, their lifestyle, their standard of living and differences for social things, they have different questions are asked.

And I have just taken here out of this different 70 questions, I have this taken around 4 questions to bring give you an example of a typical dataset. So, here if we might have seen the different questions maybe the age of the respondent then the gender of the respondent and how happy is the respondent happy maybe I have given him the 3 index maybe he is not happy, pretty happy, very happy. So, in this index are may not happy maybe 1, pretty happy maybe 2, very happy maybe 3.

Then TV hours means how many hours the respondent watch TV, so basis. So, if I consider this data, this table of data is has many numbers like now, from this number, what we can conclude? It is very difficult to conclude anything from just if we look at this table.

(Refer Slide Time: 06:04)

The slide features a title in red text: "Can You Interpret Anything from this Table". Below the title, there is a list of four questions, each preceded by a blue circular icon with a white question mark. The questions are:

- More useable information and specific answers can be obtained by organizing/summarizing the data
- What can we say about the overall frequency of the various levels of happiness?
- Do some respondents watch a lot of TV?
- Is there a relationship between age and general happiness?
- Is there a relationship between age and the number of hours of TV watched?

In the bottom right corner of the slide, there is a small video inset showing a woman with glasses and a yellow top, who appears to be the presenter. The slide also includes several decorative icons: gears, a lightbulb, a brain, and a network diagram. At the bottom left, there are logos for "Maulana Samra" and "FUTURE".

So, however, if we can somehow summarize this data, summarized this data from this table, then we can get many useful information like we can, what can we say that the overall frequency of the various level of happiness, Do some respondents watch a lot of TV these are some information you might get from the data if you can, able to summarize this data. Like is there a

relationship between age and general happiness is more or lesser age happiness is more is this such a relation is there or is there the relation between age and the number of TV hours?

So, for this, we need a data organization tools. So, we will be discussing some data organization tools, how we can summarize the data from a table of data.

(Refer Slide Time: 06:49)

Population and Sample

Population

- A population is a data set representing the entire entity of interest.
- There can be many different definitions of populations that involve the same collection of individuals.
 - The number of school-age children per household as listed in the census data would constitute a population for another study.

Sample

- A sample is a data set consisting of a portion of a population.
- Normally a sample is obtained in such a way as to be representative of the population.

Now, before summarizing, before going to the data organization tool first we need to know few other terms also like what is a population? Population, is like all of us we know what is the population it is something that is school children also know but still for completeness, let us go through it what is the population a population is a data set representing the entire entity of interest, like we know census data. So, however, there can be different definition of populations that involve the same collection of individuals.

Like from the census data day one population may be the number of school aged children per household, whatever we got in the census data, it constitutes a population for another study again the number of say unemployed youth per household that again may constitute a population for another study from the same census data we might get different definition of populations. Now, then coming to sample what is a sample?

Sample is a data set consisting of a portion of a population, but portion of a population means not any portion will just extract some portion and we get that is a sample no, that is not a sample. Normally, a sample is obtained in such a way so as to be representative of the population. Right, when I talk about the representative of the population, like let me give an example of exit poll what we do in exit poll?

So, suppose an election in an election if we were interested which party will be winning this election how we do? We definitely do not try to find out from each and every person who has come to cast their vote, we do not go and ask each and every one of them by. So, be the person who whoever some news media, whoever does the exit poll day just select a particular some person and from there try to get the information from them and that is the result of the exit poll from that day trying to infer on the whole election.

So, now, when they select these people, how do they select? The selected representative of the population when they select a representative meaning, representative meaning means they might select people from different age category they might seem select people from different profession, they might select people from different agendas mayors, and then they might select people from different area different locality this way, so, there may be so as they are the represents the whole population. Now, what is this statistical method?

(Refer Slide Time: 09:14)

What is Statistical Method?

- ❖ Method by which characteristics of a population are inferred through observations made in a representative sample from that population.
- ❖ Statistical methods are designed to contribute to the process of making scientific judgments in the face of uncertainty and variation.

Example

- ❑ Analyze the quality of a manufacturing process
- ❑ Conclusion about efficacy of a new drug

The slide features a background with faint icons of a gear, a brain, and a molecular structure. On the right side, there is an illustration of a person sitting at a desk with a laptop, and another person standing next to a large screen displaying various data charts and graphs. The slide is part of a presentation, as indicated by the navigation icons on the right edge.

Mouliya Samra
IIT KHARAGPUR

So, when we talk of statistical method, method by which characteristics of a population are inferred through observation made in a representative sample from that population. So, the exit poll that was a very good example, like what we have done, we have taken the views of a representation of some people, from their views, we try to infer for the populations. So, method by which characteristics of a population are inferred so, method by which characteristics of a populations are inferred through observation made in a representative samples.

So, whatever observation we have made in the people which we have interviewed from there we are trying to infer for the whole populations. So, that is statistical method. Again and let me first give you some other example, like analyzing the quality of a manufacturing process if you are considering the manufacturing process, suppose, this manufacturing process does it has come out it manufacture in some batches.

So, when you consider the batches the material density is producing some material maybe the density may be different vary from batch to batch even while the some it produces or in continuous then also there may be differences in the material density from different items from each item. So, we need to analyze how, how much it is different from the target density. And then we have engineer should be able to realize where the process needs improvement so, as to bring the material to close to the target density.

One another good example, conclusion about the efficacy of a new drug, let us take an example of a like a drug for curing headache suppose there is an old drug. So, which resists which gives 80% of the patients gets cured headache gets cured with this old drug and maybe this old drug we get at a cheaper price now, a new drug has come to the market. So, which it is a claim that it gives a success rate is 90%. However, the price is a bit higher.

So, what before bringing into market we need to check whether it really gives 90% of the success whether it really can achieve so, what it does, so, it has to test on few individually, it has to test on few subjects to find out whether it is really 90% efficient. Now, it has taken it has tested on a few subjects and it has found that it is 90% effective. Again suppose it has done the test again on

different set of subjects and it has come there suppose it is found, it is only 75% effective. So, there is some variance in that we got it the difference in effectiveness from study to study.

So, that is what come to the second point statistical methods are designed to contribute to the process of making scientific judgment in the face of uncertainty and variation. So, different study there may be uncertainty there may be variation. So, statistical methods are design to contribute to the process of making scientific judgment in the face of this uncertainty and variation. So, that is how we define statistical methods.

(Refer Slide Time: 12:27)

The slide is titled "Observations and Variables". It contains two text boxes: a blue one for "Observation" and a yellow one for "Variable". To the right is a table with columns "Obs", "Length", and "Weight". A bracket on the left of the table groups the rows as "Observations", and a bracket above the columns groups them as "Variable".

Observation
A data set is composed of information from a set of units. Information from a unit is known as an observation.

Variable
An observation consists of one or more pieces of information about the unit; these are called variables.

| Obs | Length | Weight |
|-----|--------|--------|
| 1 | 32 | 290 |
| 2 | 33 | 296 |
| 3 | 34 | 299 |
| 4 | 34 | 300 |
| 5 | 35 | 305 |
| 6 | 40 | 307 |
| 7 | 45 | 311 |
| 8 | 45 | 315 |
| 9 | 49 | 325 |
| 10 | 48 | 330 |
| 11 | 53 | 340 |
| 12 | 58 | 355 |
| 13 | 53 | 357 |
| 14 | 57 | 359 |
| 15 | 59 | 361 |

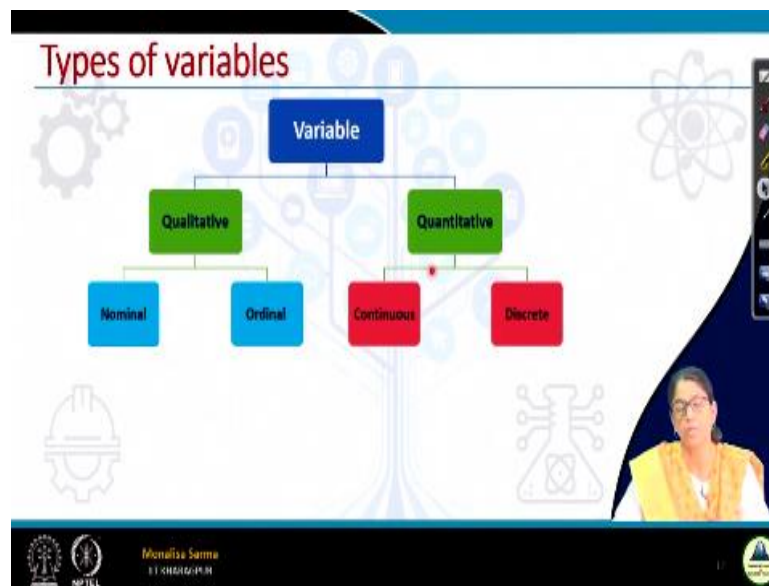
Now, some more definitions, like first is what is an observation? So, a data set is composed of information what we have already seen right from a set of units, now information from a unit is known an observation. So, now, let us take the example of these drugs that cures headache. So, what are the units in suppose we have interviewed 20 persons, so 20 total there are 20 units. So, 20 units, so 20 units means to how many observations? So, there are 20 observations each unit one observation.

So, now, next is that variable and observation consists of one or more pieces of information about a unit these are called variables. So, like this example, the drug for headache, what are the variables here variables may be the 2 variables, one is the drug use and another is the maybe the

whether the headache is cured or not, but a headache. So, this may be the second what to say variable. So, here there are total 20 which means we have interviewed 20 persons.

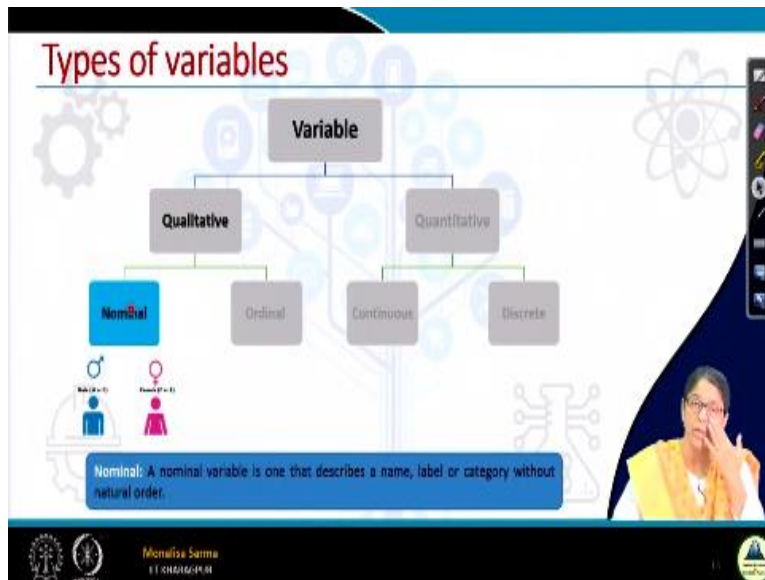
So, we got 20 observation and each observation there how many variables, there are 2 variables variable one is the drug use and another is the whether the drug is successfully using the patient determines whether a patient is cured or not cured, maybe the answer may be yes or no. So, there are 2 variables. So, similar here the example is given there are 2 variables this length is one variable weight is one variable. So, there are total say 15 observations here. So, each unit leads to one observation.

(Refer Slide Time: 14:02)



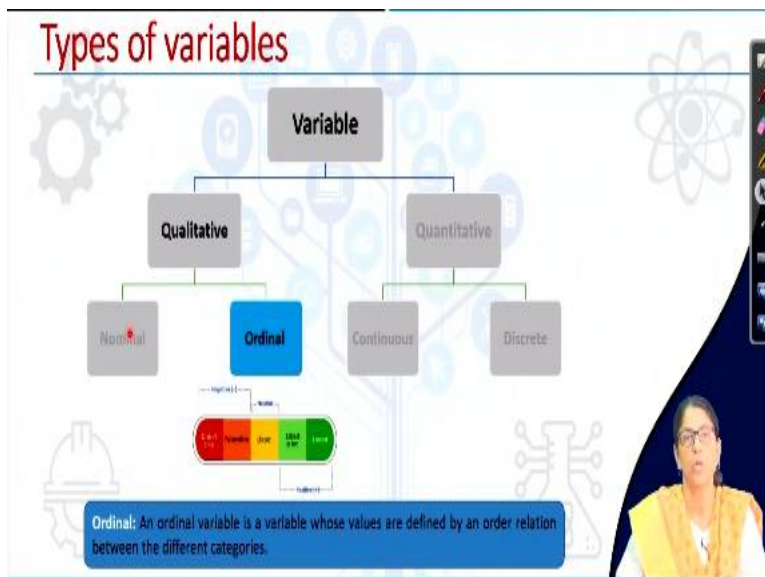
Now, this variables when you talk of variables there are different types of variables, there will be 2 types of variables one is qualitative variables it is also called categorical variable and another is quantitative variables. So, qualitative variable that is the categorical variables again there are 2 types of qualitative variables that is nominal variable and ordinal variable.

(Refer Slide Time: 14:27)



So, what is nominal variable? So, a nominal variable is one that describes a name, label or category without natural ordering like breed of animals, colours, different brands these are nominal variables.

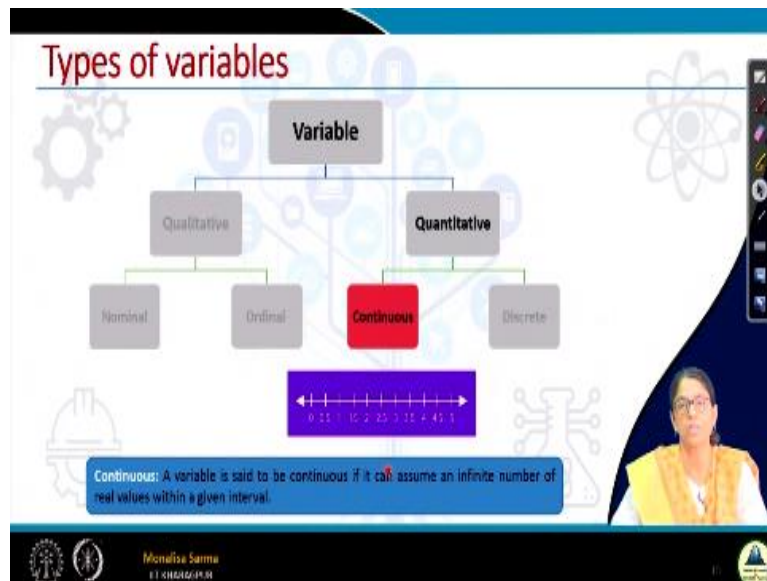
(Refer Slide Time: 14:43)



So, next is the ordinal variable, ordinal variable is a variable whose values are defined by an order relation between the different categories. Now, different categories like let me give you an example. Suppose I have asked one person, I have given him 5 different chocolate pies and I have asked him to do it in a scale of how much he has liked the chocolate pies I have given in a scale of 5 maybe he did not like it at all, he had like it bit like it he liked it very much.

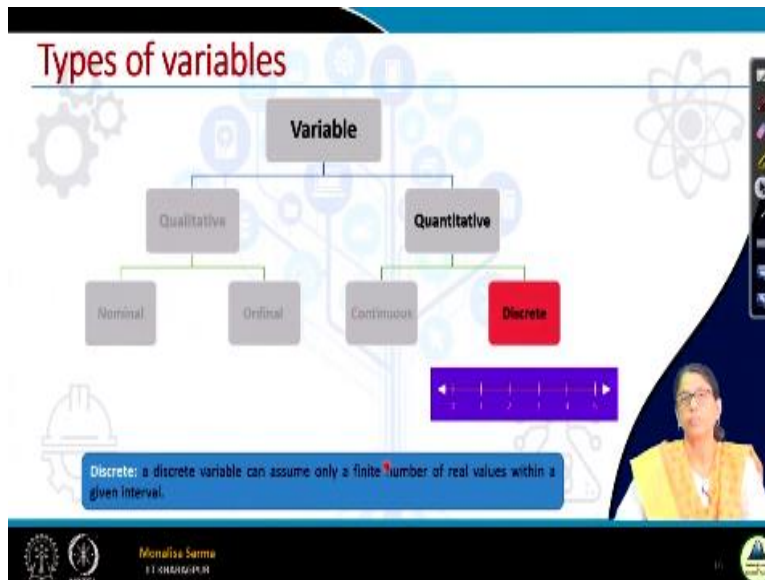
And it is excellent that maybe I have given 5 different scale and I have given him a 5 chocolate pies and accordingly rate this 5 chocolate pies in these 5 scales. So, what is this? There is a relation among the order it among the different categories. So, this is ordinal variable. In the first example, where does I have talked about a social science survey where we have asked 70 questions one of the questions is the happiness index. So, how happy he is? Is not happy, pretty happy, quite happy. So that is again ordinal variable.

(Refer Slide Time: 15:47)



So, what is a continuous variable, continuous variable all of us of course, as we know it, so, a variable is said to be continuous if it can assume an infinite number of real values within a given interval, some example of continuous variable height of a person is a continuous variable, this car mileage is continuous variables.

(Refer Slide Time: 16:08)



Discrete variable, a discrete variable can assume only a finite number real values within a given interval. So, discrete variables like a number of persons in a family in our census data number of person in a family that is a discrete variable.

(Refer Slide Time: 16:22)

Types of Variables

| | Nominal | Ordinal | Discrete | continuous | | | | | | | |
|------|---------|---------|----------|------------|------|--------|-------|--------|----|-------|--------------------|
| Obs. | Zip | Age | Bed | Bath | Size | Lot | Style | Garage | Fy | price | Rating (1-5 scale) |
| 1 | 3 | 21 | 3 | 3.0 | 951 | 64904 | Other | 0 | 0 | 30000 | 3 |
| 2 | 3 | 21 | 3 | 2.0 | 1036 | 217800 | Frame | 0 | 0 | 39900 | 3 |
| 3 | 4 | 7 | 1 | 1.0 | 676 | 54450 | Other | 2 | 0 | 46500 | 2 |
| 4 | 3 | 6 | 3 | 2.0 | 1456 | 51836 | Other | 0 | 0 | 48500 | 2 |
| 5 | 1 | 51 | 3 | 1.0 | 1186 | 10857 | Other | 1 | 1 | 51500 | 1 |
| 6 | 2 | 19 | 3 | 2.0 | 1456 | | Frame | 0 | 0 | 56900 | 2 |
| 7 | 3 | 8 | 3 | 2.0 | 1368 | 11016 | Frame | 0 | 0 | 59900 | 4 |
| 8 | 4 | 27 | 3 | 1.0 | 994 | 6259 | Frame | 1 | 0 | 62500 | 5 |
| 9 | 1 | 51 | 2 | 1.0 | 1176 | 11348 | Frame | 1 | 0 | 62500 | 5 |
| 10 | 3 | 1 | 3 | 1.0 | 1216 | 25450 | Other | 0 | 1 | 65500 | 4 |
| 11 | 4 | 32 | 3 | 2.0 | 1410 | 40057 | Brick | 0 | 0 | 69000 | 1 |
| 12 | 1 | 22 | 1 | 1.0 | 1500 | 80000 | Brick | 2 | 0 | 85000 | 1 |

Monalisa Sarma
ET CHANDIGARH

Here I have again explained the type of variables I have taken another one example, suppose this is a data of some houses which are sold in an area. In an area that is on the last set and yes, there are different houses that are sold totally you can see there are 12 observations. So, different houses are basically the same belongs to different zip code. So, we have different zip code 1 2 3 4 to maybe whatever they are, we have some zip codes and what is the age of the house? How old is the house?

And how many bedrooms are there, how many bathrooms are there, what is the size of the house? what is the lot means what is the overall area of the property, then what is how the exterior is designed, then how many garages are there fireplace, price and ratings, when people have sell, give people has given some rating to the houses. So, this may be the different these are the different variables of these observations. To total 12 observation and how many variables zip is 1 2 3 4 5 6 7 8 9 10 11 total data 11 variables now we will see which is which variable.

(Refer Slide Time: 17:30)

Types of Variables

| Obs. | Zip | Bed | Bath | Size | lot | Other | Garage | Fj | price | Ratings (1-5 scale) | |
|------|-----|-----|------|------|------|--------|--------|----|-------|---------------------|---|
| 1 | 3 | 21 | 3 | 3.0 | 951 | 64904 | Other | 0 | 0 | 30000 | 3 |
| 2 | 3 | 21 | 3 | 2.0 | 1036 | 217800 | Frame | 0 | 0 | 39900 | 3 |
| 3 | 4 | 7 | 1 | 1.0 | 676 | 54410 | Other | 2 | 0 | 46500 | 2 |
| 4 | 3 | 6 | 3 | 2.0 | 1456 | 51836 | Other | 0 | 0 | 48500 | 2 |
| 5 | 1 | 51 | 3 | 1.0 | 1186 | 10857 | Other | 1 | 1 | 51500 | 1 |
| 6 | 2 | 19 | 3 | 2.0 | 1456 | | Frame | 0 | 0 | 56900 | 2 |
| 7 | 3 | 8 | 3 | 2.0 | 1368 | 11016 | Frame | 0 | 0 | 59900 | 4 |
| 8 | 4 | 27 | 3 | 1.0 | 994 | 6259 | Frame | 1 | 0 | 62500 | 5 |
| 9 | 1 | 51 | 2 | 1.0 | 1176 | 11348 | Frame | 1 | 0 | 62500 | 5 |
| 10 | 3 | 1 | 3 | 2.0 | 1216 | 25450 | Other | 0 | 1 | 65500 | 4 |
| 11 | 4 | 32 | 3 | 2.0 | 1410 | 40057 | Brick | 0 | 0 | 69000 | 1 |
| 12 | 2 | 22 | 2 | 3.0 | 1500 | 80900 | Brick | 2 | 0 | 85000 | 1 |

So, here zip is a nominal variable, right? There is a it is defines a different given different classes which are not it has it does not have any relation as such. So, then similarly external, it is also a nominal variable.

(Refer Slide Time: 17:45)

Types of Variables

Continuous **Ordinal** Discrete Continuous

| Obs. | Zip | bed | bath | Size | lat | Style | Garage | Fa | price | Rating (1-5 scale) | |
|------|-----|-----|------|------|------|--------|--------|----|-------|--------------------|---|
| 1 | 3 | 21 | 3 | 3.0 | 951 | 64904 | Other | 0 | 0 | 30000 | 3 |
| 2 | 3 | 21 | 3 | 2.0 | 1036 | 217800 | Frame | 0 | 0 | 39900 | 3 |
| 3 | 4 | 7 | 1 | 1.0 | 676 | 54430 | Other | 2 | 0 | 46500 | 2 |
| 4 | 3 | 6 | 3 | 2.0 | 1456 | 51836 | Other | 0 | 0 | 48500 | 3 |
| 5 | 1 | 51 | 3 | 1.0 | 1186 | 10857 | Other | 1 | 1 | 51500 | 1 |
| 6 | 2 | 19 | 3 | 2.0 | 1456 | | Frame | 0 | 0 | 56900 | 2 |
| 7 | 3 | 8 | 3 | 2.0 | 1368 | 11016 | Frame | 0 | 0 | 59900 | 4 |
| 8 | 4 | 27 | 3 | 1.0 | 994 | 6259 | Frame | 1 | 0 | 62500 | 5 |
| 9 | 1 | 51 | 2 | 1.0 | 1176 | 11348 | Frame | 1 | 0 | 62500 | 5 |
| 10 | 3 | 1 | 3 | 2.0 | 1216 | 25450 | Other | 0 | 1 | 65500 | 4 |
| 11 | 4 | 32 | 3 | 2.0 | 1410 | 40057 | Brick | 0 | 0 | 69000 | 1 |
| 12 | 2 | 22 | 1 | 3.0 | 1500 | 60000 | Brick | 2 | 0 | 85000 | 1 |

Monalisa Sarma
ET ORIGINATOR

Then ratings that is an ordinal variable, there is some relative ratings is not it. So, that is an ordinal variable.

(Refer Slide Time: 17:53)

Types of Variables

Continuous Ordinal **Discrete** Continuous

| Obs. | Zip | bed | bath | Size | lat | Style | Garage | Fa | price | Rating (1-5 scale) | |
|------|-----|-----|------|------|------|--------|--------|----|-------|--------------------|---|
| 1 | 3 | 21 | 3 | 3.0 | 951 | 64904 | Other | 0 | 0 | 30000 | 3 |
| 2 | 3 | 21 | 3 | 2.0 | 1036 | 217800 | Frame | 0 | 0 | 39900 | 3 |
| 3 | 4 | 7 | 1 | 1.0 | 676 | 54430 | Other | 2 | 0 | 46500 | 2 |
| 4 | 3 | 6 | 3 | 2.0 | 1456 | 51836 | Other | 0 | 0 | 48500 | 2 |
| 5 | 1 | 51 | 3 | 1.0 | 1186 | 10857 | Other | 1 | 1 | 51500 | 1 |
| 6 | 2 | 19 | 3 | 2.0 | 1456 | | Frame | 0 | 0 | 56900 | 2 |
| 7 | 3 | 8 | 3 | 2.0 | 1368 | 11016 | Frame | 0 | 0 | 59900 | 4 |
| 8 | 4 | 27 | 3 | 1.0 | 994 | 6259 | Frame | 1 | 0 | 62500 | 5 |
| 9 | 1 | 51 | 2 | 1.0 | 1176 | 11348 | Frame | 1 | 0 | 62500 | 5 |
| 10 | 3 | 1 | 3 | 2.0 | 1216 | 25450 | Other | 0 | 1 | 65500 | 4 |
| 11 | 4 | 32 | 3 | 2.0 | 1410 | 40057 | Brick | 0 | 0 | 69000 | 1 |
| 12 | 2 | 22 | 1 | 3.0 | 1500 | 60000 | Brick | 2 | 0 | 85000 | 1 |

Monalisa Sarma
ET ORIGINATOR

Then the number of bedrooms, number of bathrooms, garage, fireplaces are discrete variable this connect the continuous variables.

(Refer Slide Time: 18:00)

Types of Variables

| QNo. | Zip | Age | Bed | Bath | Size | Lot | Other | Garage | Fa | price | Rating (1-5 scale) |
|------|-----|-----|-----|------|------|-------|-------|--------|----|-------|--------------------|
| 1 | 3 | 25 | 3 | 3.0 | 1500 | 12000 | Other | 0 | 0 | 25000 | 3 |
| 2 | 3 | 35 | 3 | 2.0 | 1800 | 11000 | Frame | 0 | 0 | 30000 | 3 |
| 3 | 4 | 28 | 1 | 1.0 | 1200 | 8000 | Other | 2 | 0 | 15000 | 2 |
| 4 | 3 | 30 | 3 | 2.0 | 1600 | 10000 | Other | 0 | 0 | 20000 | 2 |
| 5 | 1 | 22 | 3 | 1.0 | 1400 | 9000 | Other | 1 | 1 | 18000 | 1 |
| 6 | 2 | 32 | 3 | 2.0 | 1700 | 11000 | Frame | 0 | 0 | 22000 | 2 |
| 7 | 3 | 28 | 3 | 2.0 | 1500 | 10000 | Frame | 0 | 0 | 28000 | 4 |
| 8 | 4 | 35 | 3 | 1.0 | 1300 | 9000 | Frame | 1 | 0 | 35000 | 5 |
| 9 | 1 | 25 | 2 | 1.0 | 1100 | 8000 | Frame | 1 | 0 | 20000 | 5 |
| 10 | 3 | 30 | 3 | 2.0 | 1600 | 10000 | Other | 0 | 1 | 25000 | 4 |
| 11 | 4 | 28 | 3 | 2.0 | 1500 | 9000 | Brick | 0 | 0 | 18000 | 1 |
| 12 | 2 | 32 | 3 | 3.0 | 1700 | 11000 | Brick | 2 | 0 | 30000 | 1 |

Than the size, lot size, price all this is these are all these are continuous variables. Now, if we see this type of table, as I have already told before this type of table we hardly can get any information like at the most the information that we can get is the price of the house may range from particularly 30000 to say 85000 or the houses maybe in the zip code of say 1 to 5 or something like that, not much information we can gather if this data is are given in this sort of big tables. So, we need to organize this data in some form or the other for this have we needed data organizations tool.

(Refer Slide Time: 18:40)

Data Organization Tool

- Frequency Distribution
 - Frequency distributions is constructed by grouping the data into categories
 - A frequency distribution is a listing of frequencies of all categories of the observed values of a variable.

So, one of the data organization tool is the frequency distribution. So, what is frequency distribution? Frequency distribution is constructed by grouping the data into categories, we can

group the data into categories like how what are the what may be the categories, like we have seen the different zip codes where houses belonging to different area different areas, different zip codes. So, we are all different, we can if we can group the houses into different zip codes.

That is we are grouping the data into different categories based on the zip code. So, a frequency distribution is a listing of frequencies of all categories of the observed values of a variable.

(Refer Slide Time: 19:19)

Relative Frequency Distribution

- A relative frequency distribution consists of the relative frequencies, or proportions (percentages), of observations belonging to each category.
- The relative frequencies have a useful interpretation:
 - They give the chance or probability of getting an observation from each category in a blind or random draw.
 - For this reason a relative frequency distribution is often referred to as an **observed or empirical probability distribution**.

Moulija Sarma

Then we have something already relative frequency distribution, relative frequency distribution is like it is nothing but it consists of relative frequency or the proportion, proportion of observation belonging to a each category like relative frequency distribution again, it can we can also call it as in terms of percentage, a relative frequency distribution again, it has some useful interpretation also. So see, they give the chance or probability of getting an observation from each category in a blind or random draw.

So, for this reason, a relative frequency distribution is often referred to as an observed or empirical probability distribution.

(Refer Slide Time: 20:03)

Distribution of Zip

| Obs | Zip |
|-----|-----|
| 1 | 3 |
| 2 | 3 |
| 3 | 4 |
| 4 | 3 |
| 5 | 1 |
| 6 | 2 |
| 7 | 3 |
| 8 | 4 |
| 9 | 1 |
| 10 | 3 |
| 11 | 4 |
| 12 | 2 |

| zip | Frequency | Percent | Cumulative frequency | Cumulative percent |
|-----|-----------|---------|----------------------|--------------------|
| 1 | 2 | 16.6 | 2 | 16.6 |
| 2 | 2 | 16.6 | 4 | 33.2 |
| 3 | 5 | 41.8 | 9 | 75.0 |
| 4 | 3 | 25.0 | 12 | 100.0 |

So, there is an example of how we can categorize based on this zip code what I have just told. So, these are the different observations. So, based on different zip code, we have found out total zip in zip code 1 suppose there are 2 houses, zip code 2 there are 2 houses zip code 3 there 5 similarly 4 there are 3 houses, so, this is one set of category and this next is the percent that means that the relative frequency that is the proportion 16.6% of the houses total houses that are sold 16.6% of the houses belong to zip code 1 then 41.8% belongs to the zip code 3.

Then we have called cumulative frequency, cumulative frequency the first column you can see this, so, 2 so, what this second let us see the second column cumulative frequencies 4 that means, number of houses that belongs to zip code 2 and 1 together that is 4 that is the cumulative frequency. Similarly, cumulative percentage that is cumulative of relative frequency.

(Refer Slide Time: 21:04)

Graphical Representation of Distributions

- Bar chart
- Histogram
- Pie chart
- Box and Whisker plot
- Scatter plot

Monalisa Sarma
IIT KHARAGPUR

So fine, now a picture is equal to 1000 words and if we in fact instead of writing something, if a figure is given to us, it is more easier to grasp it is more easier to understand the things so, like this data organizations, if instead of just using it in some tables, if we put it in some sort of figure, some sort of graph, it is more easy to understand. So, we have some graphical representation of this distribution as well as distribution I mean the frequency distribution.

So, what the different graphical representation may be bar chart histogram pie chart, box and whisker plot, scatter plot there are some other graphical representation as well. So, we will be just discussing this.

(Refer Slide Time: 21:47)

Bar Chart (for exter)

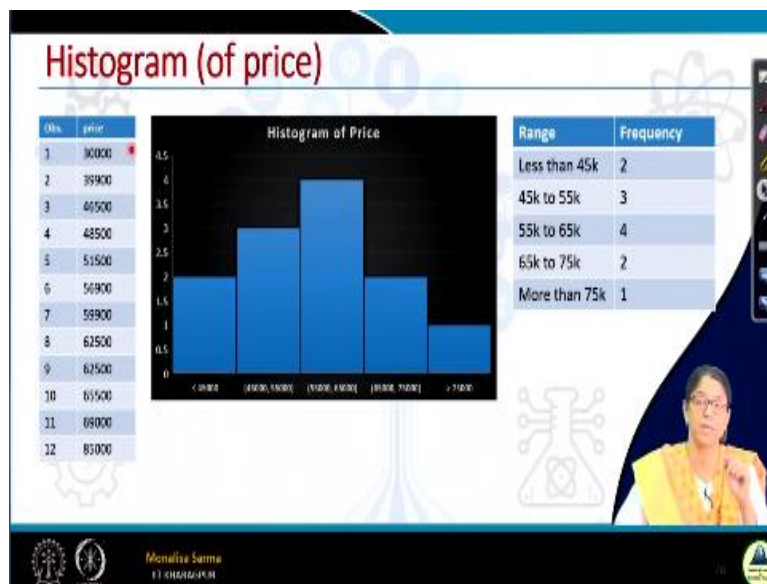
| Q.No. | Exter |
|-------|-------|
| 1 | Other |
| 2 | Frame |
| 3 | Other |
| 4 | Other |
| 5 | Other |
| 6 | Frame |
| 7 | Frame |
| 8 | Frame |
| 9 | Frame |
| 10 | Other |
| 11 | Brick |
| 12 | Brick |

Monalisa Sarma
IIT KHARAGPUR

So, when we talk a bar chart bar chart is applicable only for categorical variables what I mean by categorical variables means qualitative variables like for nominal variables for ordinal variables we use bar chart. So, here suppose for the exterior we have seen when seeing the figure we have drawn the figure for exterior so, what are the different types of exterior we had that is brick we have frame and one others. So, this is how we draw the bar chart in the y axis we get the frequency and the x axis we define the class.

So, your brick one class frame is another class others is another class and in a y axis we give the frequency. Now here the width in the bar chart width has no meaning as such, but it is just for to look it beautiful. So, as the picture does not look cluttered, we usually give the weight of each class we keep it same. So, this is the bar chart.

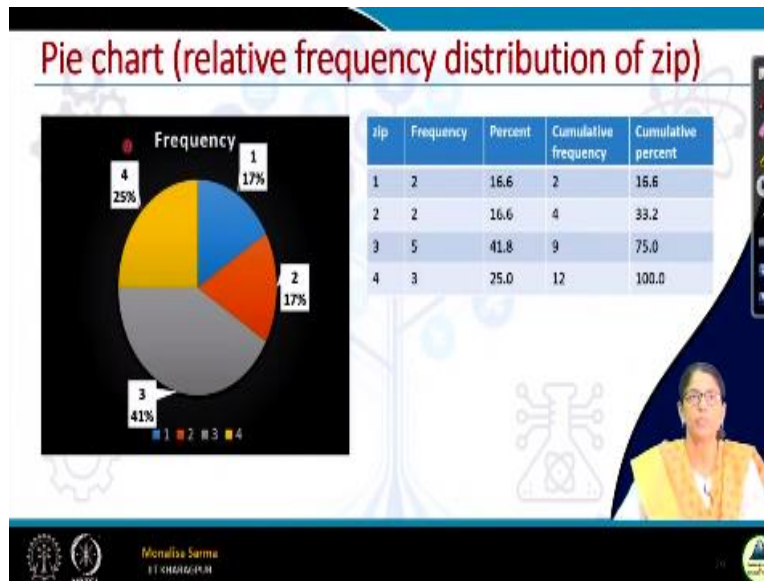
(Refer Slide Time: 22:45)



Now, if you are interested in putting in a graphical representation of this continuous data or discrete data we use histogram. Histogram is same as bar chart just that same in the sense y axis we have frequency x axis we have the different basically here x axis we have the range of data instead of class we have the range of data. So, here the width represents the range here the width has meaning. And you see here between each block there is no space because it is continuous.

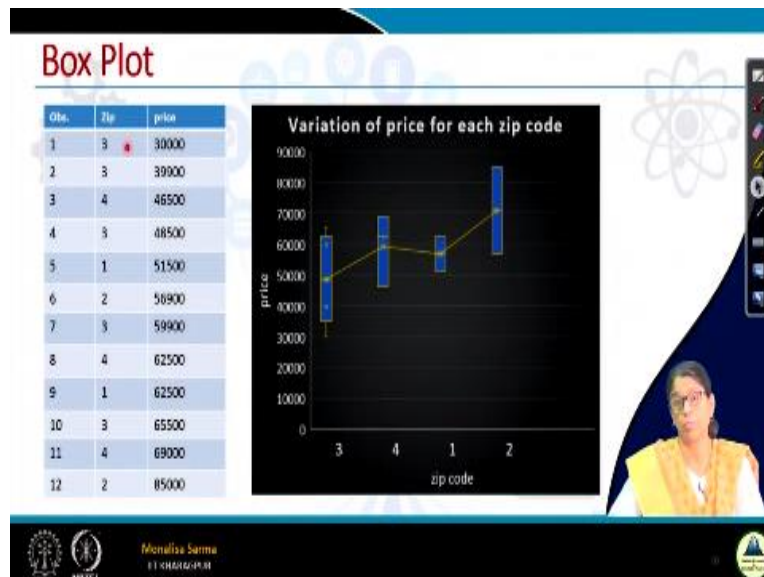
So, this is the histogram. So, we have done the histogram basically for the price histogram for the price that is for less than 45k then 45k to 55k then 55 to 65k and so on and so forth.

(Refer Slide Time: 23:36)



Then pie chart, pie chart is nothing but a circle we divide into few slices, each slice represents a proportion of the each slice represented basically category. So, category is basically the proportion, proportion of occurrence like here we have seen the for different zip code, what is the percentage of it houses are in different zip code, that is the relative proportion so pie chart, we can use pie chart to visualize.

(Refer Slide Time: 24:03)

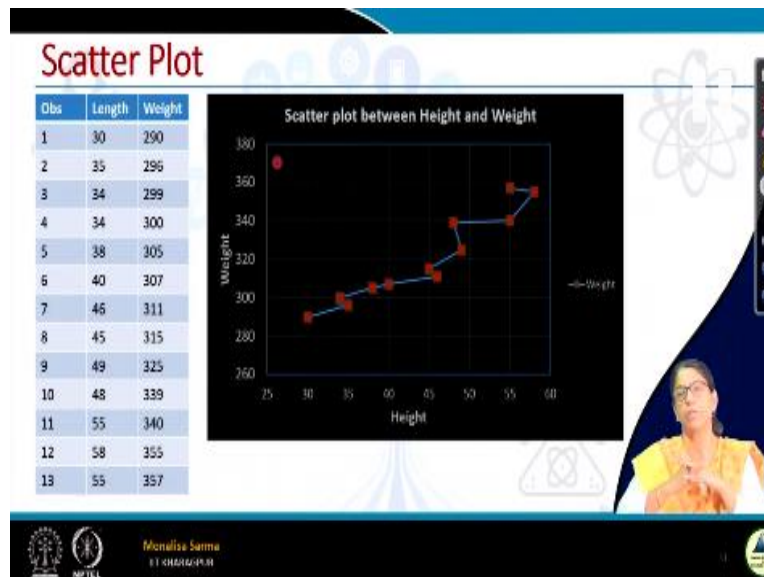


So, then we have box plot, box node is also called 5 point summary and a box plot how we draw a box plot and there are basically 3 horizontal lines, the down horizontal line is the 25th percentile, and the upper horizontal line is the 75 percentile and the middle line is the median and

then we have a whisker this is coming down, you can see a vertical line which is coming down. So, this vertical line this is coming down and here it plots the lowest value and then we have the highest value is the again a vertical line which is going up which plots the highest value.

So, here we have tried box, box plot we may use to compare different set of data and different distributions, so here we have tried to compare variation of price for each zip code for each zip code what are the variation or prices like for zip code 3 we have seen this is the variation of prices, we have drawn the box plot similarly, this is the variation of prices for box plot 4. So, when we have this sort of figure we can easily compare how the price is different from different zip code. So, this is the box plot.

(Refer Slide Time: 25:18)



Then we have scatter plot, scatter plot is like when we try to find out how one variable varies compared to other variables like how height varies compared to weight or weight varies with in accordance to height. So, this is how we plot scatter plot. So, now, okay, graphical description is okay, so again graphical description we get lots of information however some sometimes we do not need such we do not need lots of information.

We may need a simple one just one data, we may need some summary data of some certain population, we just need a summary data then we do not go for a graphical description of this

data then what we will do? We will take help of descriptive statistics to describe the characteristics of our populations okay.

This graphical description also does that it describes the characteristics of the population, but we have lots of information there, but if we want just a one value to describe the characteristics of the population this one value can be anything.

(Refer Slide Time: 26:23)

Numerical Descriptive Statistics

$X = \{x_1, x_2, x_3, \dots, x_n\}$

Mean

- The **mean** of a set of observations is the **arithmetic average** of the values.

$$\text{mean} = \frac{1}{n} \sum_{i=1}^n x_i$$

Median

- The **median** is the value separating the higher half from the lower half of a data sample.
- $\text{median} = \begin{cases} \frac{x_{(n+1)}}{2}, & \text{if } n \text{ is odd} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2}, & \text{if } n \text{ is even} \end{cases}$

Mode

- The **mode** is the value that appears most frequently in a data set.

Muralisa Sarma
11 SHS04010101

So, that the different descriptive statistics may be mean all of us so, we know what is a mean? Mean is a set of observation of a set of observation is the arithmetic average of the values all of us we know what is arithmetic mean. So, then median; median is the middle value. Now, if given a set of data, given a set of data, whether we will be considering the mean or median that is also that also it is very much subjective like if there are many outliers like let me take consider the example of a rainfall.

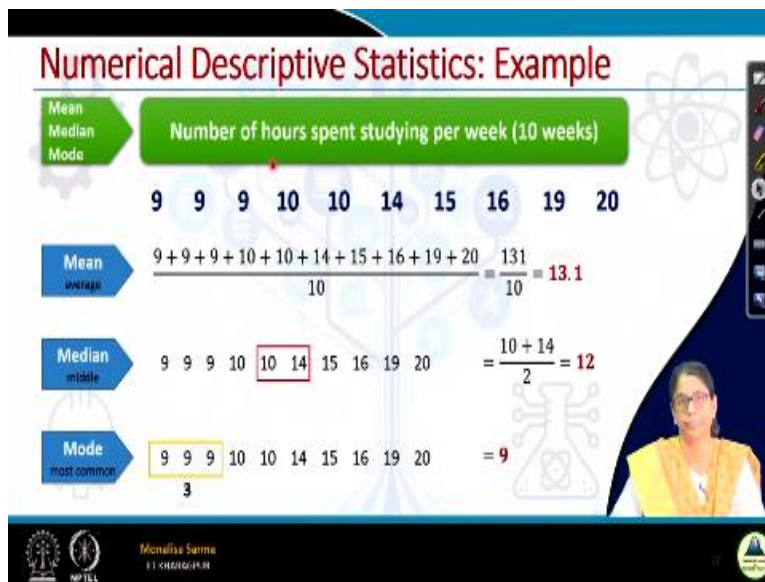
Rainfall in an area, rain fall in an area if we consider say suppose in a 30 day period 30 day period if we want to find out what is the rainfall in the area, average rainfall in an area, suppose there are lots of outlayers like sometimes there is no for somebody and then there is no rainfall at all and sometimes some days they are very huge rain fall say around to the extent of say around 120 Mm, or 200 Mm, some like something like that.

So, then if we try to find out if we try to take the mean, the mean will not be a proper reflective of the picture, because there are some very small data, some very high data. In that case, maybe we may go to the to find out the median and what is the middle data means in a 30 days period we will see in the 15th day what is the data, but then median the disadvantage with a median is that median does not take care of all the values when we consider mean, we take the average of all the values.

But when we consider median, we just take the middle value, middle value, if there is a odd number, we take the middle value, if it is even number, we take this and by 2th value and n plus 1 by 2th values, and we take the average of these 2 values. So that is the median. So that basically in which situation we will use mean in which situation you will use median that is totally very subject and dependent on the situations.

Then what is mode? Mode is the value that appears most frequently in a data set is a value that appears most frequently the data set, that appears most frequently the data set.

(Refer Slide Time: 28:24)



I will show you some example. Like suppose here is the data for number of hours spent studying per weeks and is number of hours in the given for certain weeks. That is in the number of hours a student has spent in studying. So, if I am interested in finding the mean, how do, I find the mean, I will add up all this data since it is a data for 10 weeks, then will be divided by 10. That is how I

got the mean, if I am interested in the median, so there are total 10 data sources there is an even number I will have to consider 2, middle data.

So, to the middle data that is the 10 and the 14 as a red dot the data in the red box. So that is $10 + 14 / 2$ that is my median is 12. Now if I am interested in finding the mode, mode see, your 9 has appeared more number of times so my mode is 9.

(Refer Slide Time: 29:14)

Numerical Descriptive Statistics

Variance

- Variance is a measurement of the spread between numbers in a data set.
- $variance(sample) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$

Standard Deviation

- The standard deviation of a random variable, sample, statistical population, data set, or probability distribution is the square root of its variance.
- $standard\ deviation = \sqrt{variance}$

Monalisa Sarma
II SEMESTER

Now, there is one more descriptive statistics that is variance, variance basically it gives us spread how spread out my data is how my data is different, how much spread out my data is from the mean compared to the mean how spread out the my data is compared to the mean. So, how do I calculate variance? This is the formula for calculating variance. All of you of course, that variance standard deviation mean, median, mode, you have studied in class level only.

This just a quick recap basically. So, this is variance we have calculated a variance. So, then from the variance, how we calculate the standard deviation? Standard deviation is nothing but the square root of variance.

(Refer Slide Time: 29:58)

Numerical Descriptive Statistics: Example

Variance
Stand. Dev. Number of hours spent studying per week (10 weeks)


9 9 9 10 10 14 15 16 19 20

Variance
sample

$$\frac{(9 - 13.1)^2 + (9 - 13.1)^2 + (9 - 13.1)^2 + (10 - 13.1)^2 + (10 - 13.1)^2 + (14 - 13.1)^2 + (15 - 13.1)^2 + (16 - 13.1)^2 + (19 - 13.1)^2 + (20 - 13.1)^2}{10 - 1}$$

$$= \frac{(-4.1)^2 + (-4.1)^2 + (-4.1)^2 + (-3.1)^2 + (-3.1)^2 + (0.9)^2 + (1.9)^2 + (2.9)^2 + (5.9)^2 + (6.9)^2}{9}$$

$$= \frac{16.81 + 16.81 + 16.81 + 9.61 + 9.61 + 0.81 + 3.61 + 8.41 + 34.81 + 47.61}{9}$$

$$= \frac{155.29}{9} = 17.25$$


Monalisa Sarma
IIT KHARAGPUR

So now, we can again the same example, but what we have considered for number of hours spent studying per week that is for 10 weeks. Let us take the same example for same example how do I calculate the variance?

(Refer Slide Time: 30:11)


Numerical Descriptive Statistics

Variance

- Variance is a measurement of the spread between numbers in a data set.
- $\text{variance}(\text{sample}) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$

Standard Deviation

- The standard deviation of a random variable, sample, statistical population, data set, or probability distribution is the square root of its variance.
- $\text{standard deviation} = \sqrt{\text{variance}}$



Monalisa Sarma
IIT KHARAGPUR

You see in the variance what is the definition for variance we for each data, I try to find out how much each data is different from the mean that is the \bar{x} , \bar{x} is mean. And since I may get negative values as well as positive values, so, negative values may bring out to be 0 that is why I have squared it that is you are see this a square and this is the what is the degree of freedom degree of freedom is $n - 1$. So, it is I have divided by $n - 1$ see I have not divided by n . So, since the degree of freedom is not n degrees of freedom is $n - 1$. So, this is my variance.

(Refer Slide Time: 30:46)

Numerical Descriptive Statistics: Example

Variance Stand. Dev. **Number of hours spent studying per week (10 weeks)**

9 9 9 10 10 14 15 16 19 20

Variance sample

$$\frac{(9 - 13.1)^2 + (9 - 13.1)^2 + (9 - 13.1)^2 + (10 - 13.1)^2 + (10 - 13.1)^2 + (14 - 13.1)^2 + (15 - 13.1)^2 + (16 - 13.1)^2 + (19 - 13.1)^2 + (20 - 13.1)^2}{10 - 1}$$
$$= \frac{(-4.1)^2 + (-4.1)^2 + (-4.1)^2 + (-3.1)^2 + (-3.1)^2 + (0.9)^2 + (1.9)^2 + (2.9)^2 + (5.9)^2 + (6.9)^2}{9}$$
$$= \frac{16.81 + 16.81 + 16.81 + 9.61 + 9.61 + 0.81 + 3.61 + 8.41 + 34.81 + 47.61}{9}$$
$$= \frac{155.29}{9} = 17.25$$

Monalisa Sarma
IIT KHARAGPUR

So, here so, accordingly what is my mean I have calculated mean earlier it is 13.1. So, I have substituted each data from 13.1 and divided by 10 - 1 total this 10 data so, and I got this variance now, from the variance I could calculate the standard deviation.

(Refer Slide Time: 31:04)

Conclusion

- In this lecture we got introduced to the importance of statistical learning in data analysis.
- We learned the basic concepts of data and its representations.
- We learned the concept of descriptive statistics.

Monalisa Sarma
IIT KHARAGPUR

So, in this lecture we got introduced to the importance of statistical learning and data analysis and reliability analysis whatever it is, let me tell you the general terms data analyzes this because reliability means we will be analyzing the reliability data only. That is where we will be using statistics. So, we learned the basic concept of data and its representation. We also learned the concept of descriptive statistics.

(Refer Slide Time: 31:32)



In the reference for this lecture, as I have taken the help from this book probability and statistics for engineers and scientists by Walpole, Myers, Myers and Ye is a very good book. Thank you.