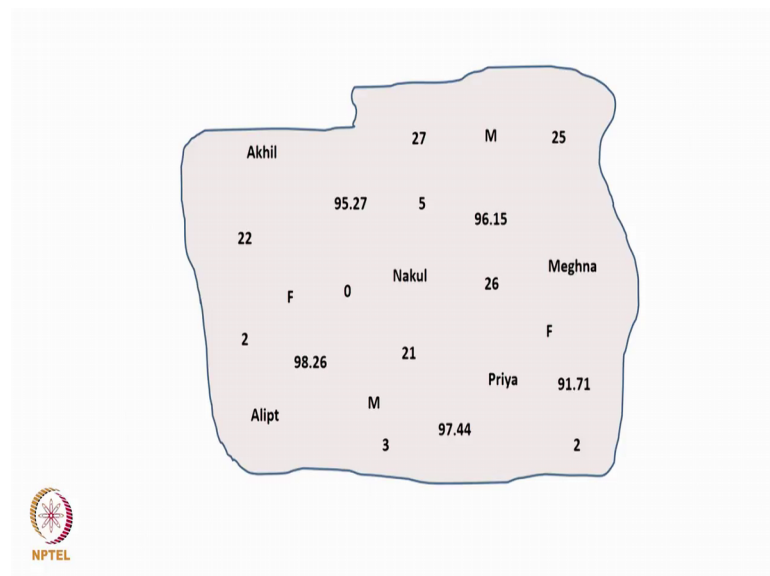


Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 02
Types of Data

In this lecture, we look at data in little more detail. We try to find out what types of data exist, and then we try to understand when we use what types of data.

(Refer Slide Time: 00:35)



So, to understand this let us just look at a small picture where we have a lot of things that are written down. And later we classify them or bring them into data. So, you could find some names here you could find some names, like Akhil or Alipt or Nakul or Priya. You could find names we could also find things like M F which kind of make us understand that they could represent male and female.

So, you find numbers like 98.26, 95.27 which might mean something which could either mean average marks. Or perhaps, they could mean some kind of a rank or a percentile or something. And then you find some numbers like 2, 3, 2 and so on. So now, all these are data and we will use this as an example to try and understand the various types of data. As well as various categories of data and so on. So, from this we can understand the data need not only mean numbers, but even names could represent data, even symbols could represent data. For example, M and F could represent male and female which are

symbols or notations to represent data whereas, we have names like Meghna or Nakul and so on.

(Refer Slide Time: 02:15)


Data table

Name	Gender	Age	Score	Experience
Akhil	M	27	95.27	5
Alipt	F	25	96.15	2
Meghna	F	26	91.71	3
Nakul	M	21	97.44	0
Priya	F	22	98.26	12

years percentile years

12 months

Columns are variables
Rows are cases or observations (n)



So, this is some piece of data picked up. And then we now bring all these data into a table. Now we kind of sort and categorize this data. So, we are able to do this we are able to for example, say that there are 5 names so these names are being written. And we could for example, associate a male female with the names which is also written. And then we look at this set of numbers which are like 27, 25, 26. And then we bunch the set of numbers which are 95.27 96.15 and so on, and then we look at 5, 2, 3, 0, 12 and so on.

I mean if you look at it very, very carefully all these 25 pieces of data may not be there in the previous picture. But one could say that this type of a table can always be drawn from the source from which the data on the previous picture had been taken. Now we have classified this data, we look at this a little more, and then we can give some generic headings as named for the first column which has Meghna, Nakul, Priya and so on. Obviously, the second one could be the gender, and that could represent male and female corresponding to the names that are there.

Now, one look at the third column one can think of many many things that the third column could represent. And perhaps, the third column could represent age of this person. So, it could represent, the age of this person assuming that these are students from a class, let us say an MBA class or whatever. So, this could kind of represent the

age of of there. The 4th one could represent a score. For example, they could represent a percentile score in an entrance exam based on which they were admitted. And the fifth could possibly represent a work experience and so on.

Even though something like 12 for a person aged 22 is inconsistent. So, they just represent some pieces of data which are there. So, broadly one could categorize this. So, the data that we have are now put in a data table, with each column having a heading, and the data fits into this heading such as name gender age score and experience. So, we also need units for some of them. So, age would be years and experience could be years while the others may not have an explicit unit in which it is measured. Score could be measured as percentile, and the 12 month is an outliers. So, I am just explaining that data could have outliers. And we need to collect and compile data carefully. So, one should also be able to understand outliers in data which is shown by the 12 months.

Now, once we make a table like this the columns are called variables, such as name gender age score and experience, and the rows are called cases or observations. It is a general terminology that is being used, the columns are called variables the rows are called cases or observations. Now what type of data are these? For example, if you look at this, one can understand that columns 1 and 2 represent data. There are not numbers whereas, columns 3, 4 and 5 represent data that are numbers.


Sometimes it is also customary to represent in this case we have M and F representing the gender. At times we could use a different notation like a 1 and 0 and so on. But columns 1 and 2 generally do not have numbers representing the data; whereas, 3 4 and 5 have numbers representing the data.

(Refer Slide Time: 06:26)

Types of data

Categorical – Responses that belong to groups or categories
Yes/No
Strongly agree to strongly disagree

Numerical - a numerical values as a response
discrete number or continuous
number of students in a class
height of people in a locality




So, how do we classify types of data? First classification is called categorical data and numerical data. So, if we go back to the first one, the table the first 2 are categorical and the next 3 are numerical. So, categorical data are responses that belong to groups or categories. They could sometimes be a yes no type of a thing. There could be something like a strongly agree to strongly disagree and so on. Numerical data uses a numerical value as a response. So, it could be a discrete number or it could be a continuous number, example number of students in a class height of people in a locality and so on.

(Refer Slide Time: 07:14)

Types of data

Qualitative – No measurable meaning to the difference of numbers
Number in the shirt of a sports person
includes nominal and ordinal

Quantitative - meaning to the difference
80 marks and 60 marks



So, first classification is categorical data and numerical data. Another classification is qualitative data and quantitative data. So, when you say qualitative data, there is no measurable meaning to the difference of numbers. For example, number in the shirt of a sports person. You could find one cricketer wearing a number 12, and another cricketer wearing a number 82. They actually do not mean much at all they just describe something.

You cannot distinct, while it helps in distinguishing say that if I see the number 12 I know this is the sports person, and I see the number 82, I see another person, but there is no way to say that the person wearing an 82 is a senior player compared to the person wearing number 12. Qualitative data are further divided into 2 types which are called nominal data and ordinal data.

We also have quantitative data which where we can give some meaning to the difference. For example, somebody has scored 80 marks and the other has scored 60 marks. Then instances one can say that this person has scored more than the other, and in some other instance one could say has scored twice the mark compared to the other. So, within the quantitative we have interval and ratio, within the qualitative we have nominal and ordinal.

(Refer Slide Time: 08:57)


Types of data

Categorical – Nominal (no implied order), Ordinal (order or rank)
 Numerical – Interval (add/subtract) , ratio (also multiply and divide)

Akhil	M	27	95.27	5
Alipt	F	25	96.15	2
Meghna	F	26	91.71	3
Nakul	M	21	97.44	0
Priya	F	22	98.26	12

nominal
nominal
ratio
ordinal
interval

Marks given for work experience



So, there are 4 broad classifications or types of data, nominal data, ordinal data, interval data and ratio data. So, categorical, nominal, no implied order, ordinal order, or rank,

numerical data classified to interval where we can add and subtract, and ratio where we can also multiply and divide in addition to add and subtract. Name is a nominal type of data. No implied order gender is nominal. So, in this case you say either male or female, qualitative data, ratio age is a ratio. So, one could say that this person is twice as old as the other so, it is a ratio type of data.

The percentile in the qualifying examination is an ordinal type of data, there is an order or a rank, one can say that somebody who got 98.26 had a higher rank than somebody who got a 97.44 at the same time we cannot say that this person has scored say one mark more, cannot say that because these are percentiles, and these only represent a rank of the marks score. So, one cannot go back and say, that the person who got 96.15 got one mark more than the person who got 95.27.

But what it represents is this person who got 96.15 is in the top 96.15 percent of those who wrote the exam. Whereas, the one who got 95.27 is within the top 95.27 of those who wrote the exam, so it is ordinal data. The work experience can be an interval data, one can go back and say that the person who has 3 years work experience has one more year work experience than the person who has 2, but it is not very fair to conclude that this person has one and half times is more work experience.

So, we now see all the so we are in nominal ordinal interval and ratio. So, we find examples of all the 4 types of data in this. So, given a certain description of data. It is very, very important for us to understand what category it comes. Most of the times have observed that it is just that bit difficult to distinguish between interval and ratio. Ordinal is reasonably all right, because you only find a rank nominal is easy relatively easy to kind of identify. Whereas, it is often difficult to distinguish between interval and ratio.


So, one needs to just understand this point very carefully that an interval we say add and subtract make sense ratio all 4 makes sense. So, the example where we said interval is while we say that the person with 3 years work experience has one more year than the person with 2, it is difficult to say that the person has one and a half times the experience or knowledge. Therefore, we categorize them as interval. So, it is important to given the type of data to quickly understand what type of these 4 it fits into, and that comes by constant practice and also by understanding the context in which the data has been picked or the data is going to be used.

(Refer Slide Time: 12:44)

The following data was collected from 100 managers :

1. Salary (range)
2. Car model
3. Year of graduation
4. Years of experience
5. Highest degree
6. Number of companies worked
7. Computer model
8. Number of countries visited
9. Number of children
10. Favourite sport

Classify the data into the four types. Give units for numerical data




For example, if we could give marks for work experience instead of using years, again one could only look at it as an interval type of a data. So, this is another example. So, this could be some kind of a class work for you. So, following data was collected from 100, managers the salary range of salary in the sense say 10 thousand to 20,000, 20,000 to 50,000.

Car model that they have, the year of graduation years of experience, highest degree number of companies that they have worked what kind of a computer they have which brand number of countries they have visited, if they are married or the number of children then they have, and what is their favorite sport. So now, you realize that there are 10 different types of data, and you could try and classify these into the 4 types that we saw nominal, ordinal, interval and ratio. And we also can give some numerical units for the numerical data. For example, one could go back experience as years and so on.

(Refer Slide Time: 13:45)

Write relevant data variables for the following situations:

1. MBA admission
2. Dental clinic
3. Savings bank
4. Automobile dealer
5. Purchase department in a factory
6. School
7. Supermarket
8. Cricketer database
9. IITM faculty profile
10. Museum



One could also go back and try to look at what kind of data I; remember in the last lecture we gave examples of data. So, similarly if you look at context like an MBA admission or a dental clinic or a savings bank or a automobile dealer or a purchase department and a factory, school, supermarket, database of cricketers, IIT madras faculty profile or a faculty profile of any educational institution, a museum. So, here we would first you can collect about 10 to 20 types of data in this. And then classify them into nominal ordinal interval and ratio.

So, with this we come to the end of the second lecture which is on data. And in this lecture we saw different types of data, and more importantly we understood the data table, and we understood that the columns are variables while rows are cases or observations. And then we went on to classify data, it is qualitative quantitative categorical numerical, and then was in the category we said nominal and ordinal as categorical and interval and ratio as numerical. And we also gave some examples to understand the characteristics of each one of them.

The most important thing being interval add and subtract there is a ratio we could add subtract multiply and divide and make meaning out of these elementary operations. In the next lecture we would look at some examples of data. And then we would also try to look at categorical data in little more detail.