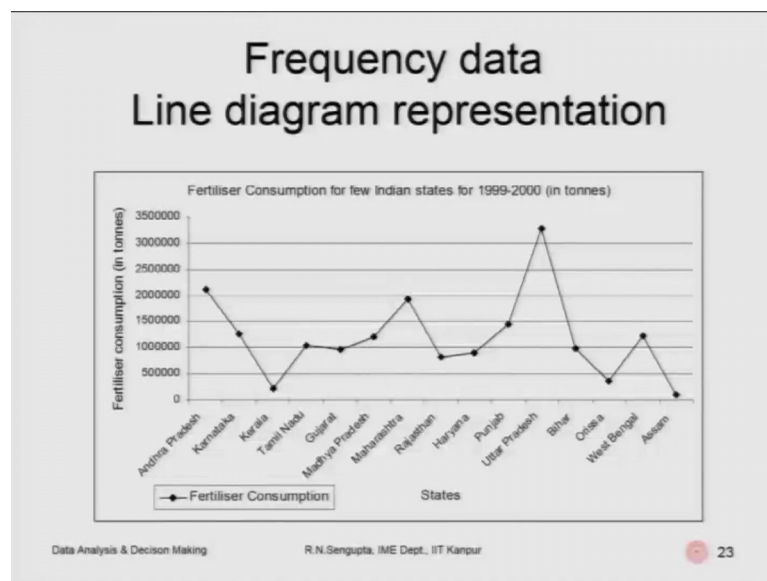


Data Analysis and Decision Making - I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 02
Data Representation & Frequency

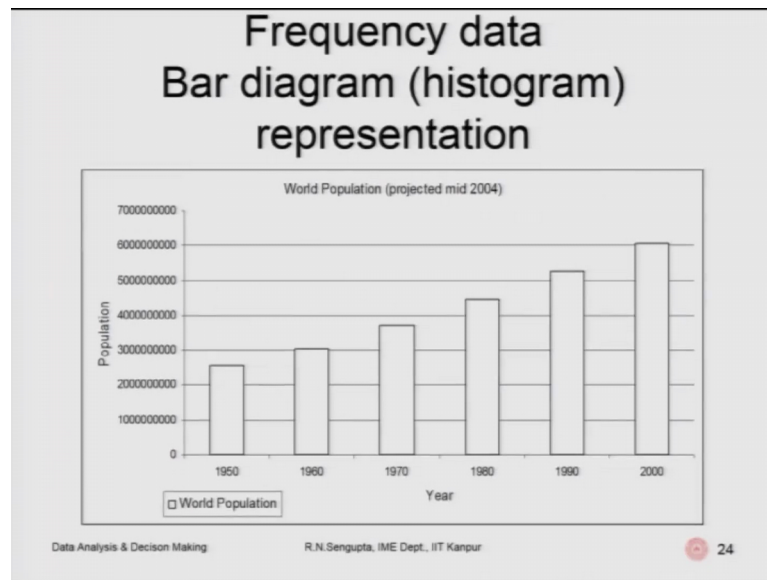
A very Good morning, Good afternoon, Good evening my dear friends. This is the second lecture as you can see on the Data Analysis and Decision Making which is basically the first part of the 3-part series for DADM. And we are considering different concepts of data representation as a frequency, and non-frequency part. And this would be the second lecture in the NPTEL MOOC series. And I am Raghu Nandan Sengupta from IME department IIT, Kanpur.

(Refer Slide Time: 00:48)



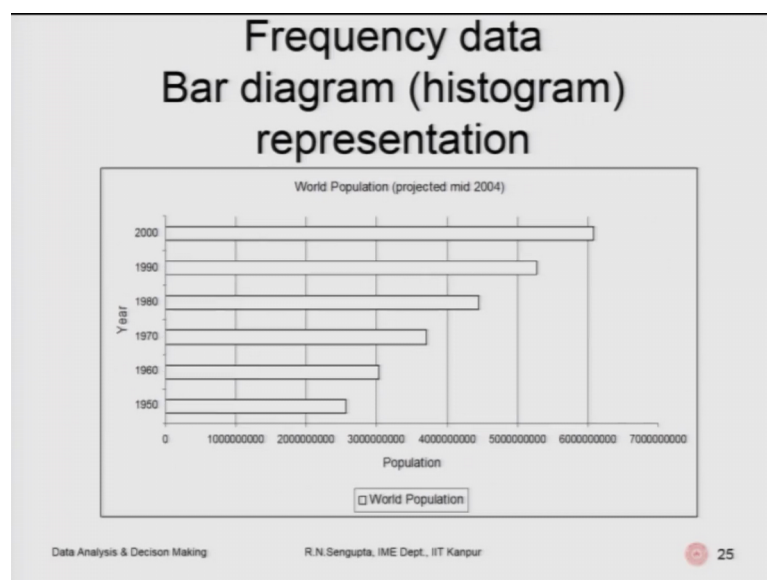
So, we are discussing the frequency of the line diagram type. So, again I am giving the same example which is the fertilizer consumption in metric tons for starting for Andhra Pradesh to Assam for the year 1999 and 2000.

(Refer Slide Time: 01:00)



Now, the frequency data diagram can representation can also be done in the bar diagram the histogram type. See for example, we have the world population starting from 1950 to 2000, and the population is given along the y-axis and the year 1950, 1960 1970, 1980, 1990 and 2000 given on the x-axis. And the height of the histogram basically gives you the value of the population or the total population.

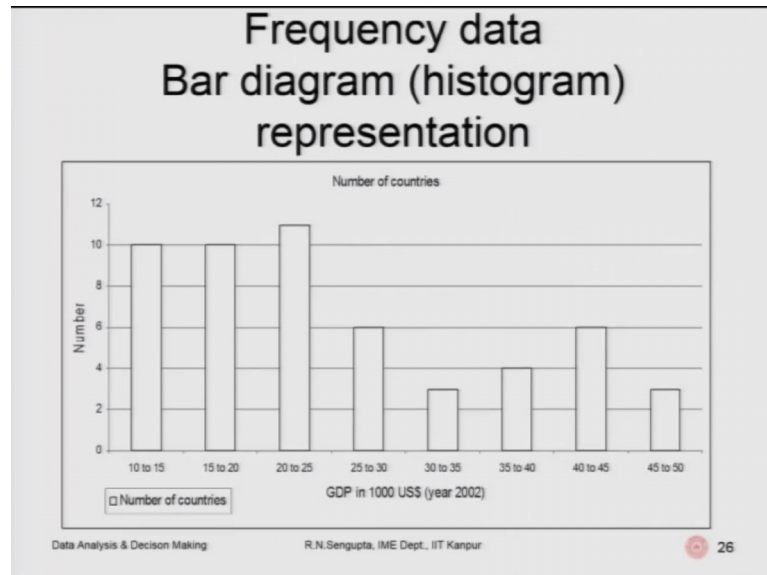
(Refer Slide Time: 01:24)



Frequency data or bar diagram can also be represented as a horizontal one or that the initial one was the vertical one, this is the horizontal one we have again. In this case, the

population values are noted on along the x-axis, while in along the y-axis you have the years starting from 1950 to 2000.

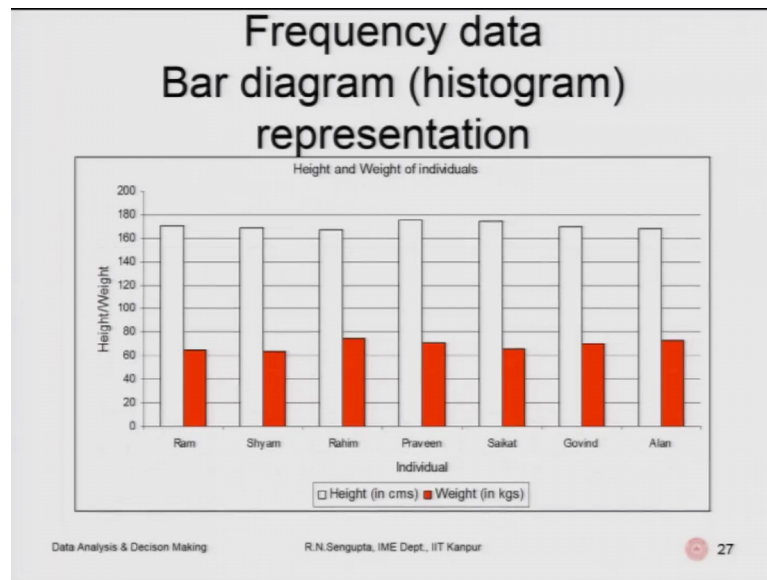
(Refer Slide Time: 01:45)



The third way of trying to describe the bar diagram in the histogram representation can be GDP in 1000 US dollars for the year 2002 for the number of countries. So, the number of countries are represented along the x-axis y-axis, and the GDP on the number of country countries as a block as a group is basically denoted along the x-axis.

So, if you see the histogram the height, if it is number of countries is 10, it basically means the GDP in 1000 between 10 to 15,000 is given for 10 countries. If we see, say for example GDP or 45 to 50 is basically for about three countries. So, if you see the height that will give you the number of countries. And along the x-axis you have the basically the grouping or the GDP in 1000 between values of say for example, 10 to 15, 15 to 20 and so on and so forth.

(Refer Slide Time: 02:39)

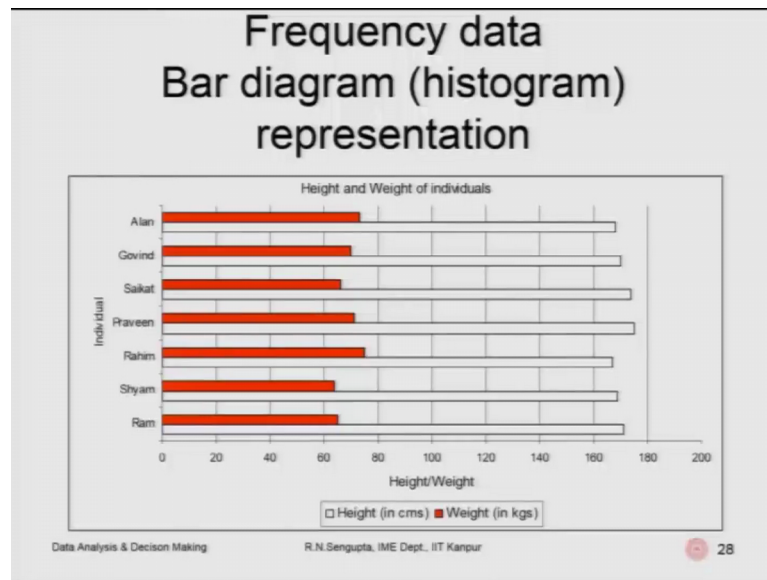


Now, if you consider the frequency diagram bar diagram, but for now we are basically slowly trying to collate or consider more than one set of information. So, consider that we have individuals Ram, Shyam, Rahim, Praveen, Sakat, Govind, and Alen. So, we have the height and the weight. So, the height in centimeters is giving along the y-axis and similarly the weight is also given in kgs along the y-axis, but there is a different color nomenclature.

So, the histogram which is white in color is for the heights, the histogram which is black the red in color is for the weights. And if you see the Ram, consider he has actually about height roughly I am giving in words is about 170 centimeters and weight of about 60 plus right say for example, 65 kgs.

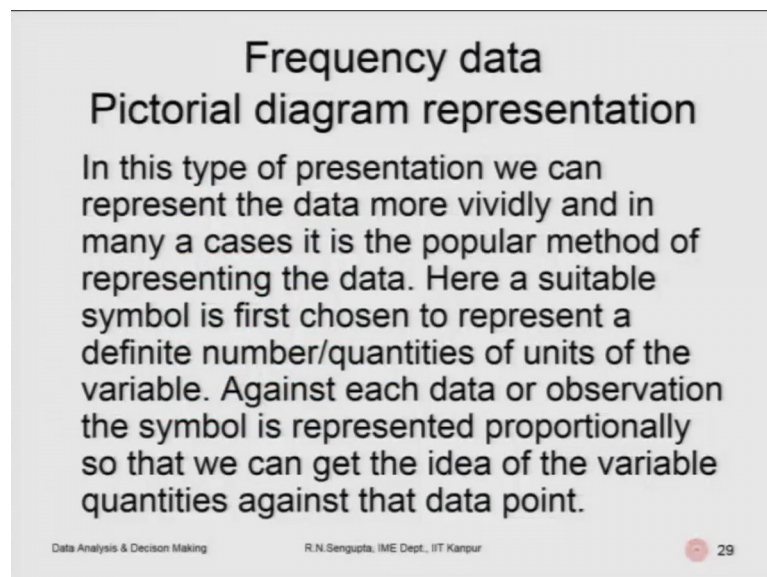
For say for example, Govind, it will be on almost the same. So, it is about 170 centimeters as height and weight would be about 70. If I consider say for example, Saikat, he is almost touching the height of 180 about 178 one 177 and his weight is about 65. So, you can find that you can have 2, 3 or more than 3 variables which you can basically depict on the histogram, and you can compare them such that you can find out the comparative study of the variables for any individuals as shown in this diagram.

(Refer Slide Time: 04:15)



So, this in the initial case, it was vertical. Now, in this case it is horizontal. We measure the same thing height and weight where the height and weight are now measured along the x-axis and the individuals are may been measured along or noted along the y-axis.

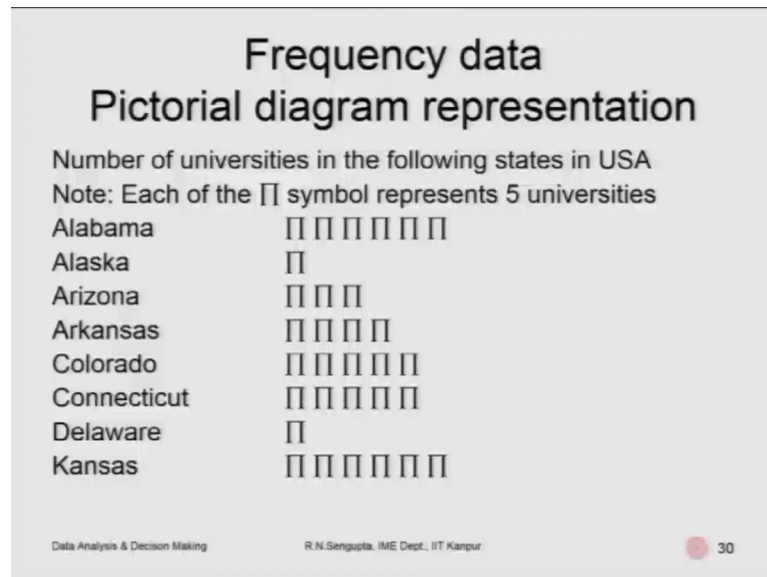
(Refer Slide Time: 04:21)



So, frequency data representation or the pictorial diagram representation, in this type of representation we can represent the data more vividly and in many a cases it is the popular method of representing the data. Here a suitable symbol is first chosen to represent a definite number or quantities of units of the variable. Against each data

observation the symbol is representative proportionally, so that we can get the idea of the variable quantities against the data point. And number of such symbols which are being utilized in whatever proportions would give you the number of times you want to basically depict the quantum of the data.

(Refer Slide Time: 05:01)



So, consider in this concept of frequency data representation which is the pictorial diagram representation. It basically denotes a number of universities in the following states in USA. So, note each symbol, all this pi symbol consider this; the symbol will represent 5 universities. So, in Alabama, you will basically have 1, 2, 3, 4, 5, 6, 6 into 5, so 30 such universities are there. If I consider Delaware, it would be about 5 universities. If I consider say for example, the state of Kansas, it will be 6 into 5 about 30 universities and so on and so forth.

(Refer Slide Time: 05:33)

Frequency data
Statistical map representation

If for example we are interested to show diagrammatically regional seismicity in Alaska of earthquakes of all magnitudes reported between 01/01/1960 to 11/09/2002. The colour code as shown indicates the depth of the event. Thus blue: $0 < h \leq 33$ km, green: $33 < h \leq 75$ km, red: $75 < h \leq 125$ km and yellow : $h > 125$ km. The larger circles are earthquakes of M 7.0 and higher from 1900-11/09/2002. The colour of circle indicates the depth of the event, as above. The star indicates the location of the 03/11/2002

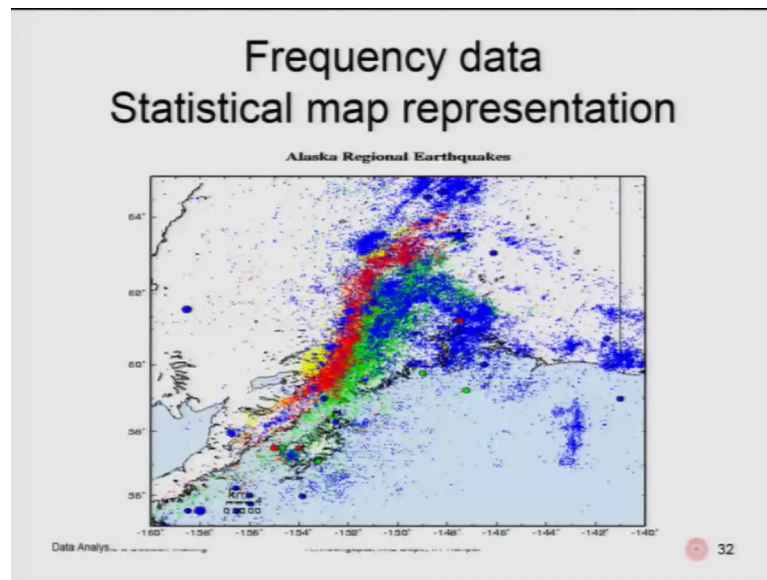
Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 31

Next is an example of frequency data a statistical map representation. If it is used for example, and let us give their example here. So, we are interested to show that diagrammatically the regional system at seismicity for earthquakes in Alaska of all magnitude which have been reported between the year 1960 to 2002. So, that means from 1st January 1960 to 11th of September 2002.

The color coding has been utilized in this sense. Thus if it is blue in color, it means the height or basically the depth at which the it has occurred and the depth of the event. Height means basically from the level where you are considering the earthquake. If it is between 0 to 33; it is given blue in color. If it is 33 to 75 kilometers, it is green in color. If it is red, it is basically 75 to 125 kilometers. And yellow it is more than 125 kilometers.

The largest circles are the circled circumference of the circle diameter of the circle area are basically given from on a Richter scale of 7, and for the values the color scheme basically denotes the circle indicates is the depth of the event, as above. And the star indicates the location as on 03.11.2002.

(Refer Slide Time: 06:56)



So, if you see the map here, this is the Alaska region. So, basically we have the you can have the latitude and longitude, the circles, and to their area specific would basically depend that how vast the effect of the earthquake have been. And the color scheme would basically depend that at what height or what depth I should not use the word height at what depth the earthquake has happened.

So, you can understand, so if it is blue in color, red in color green in color, you can understand both its intensity and at what depth the actual earthquake has happened. So, this is basically for the map of Alaska. So, obviously, you can do it for our different regions accordingly it can be rainfall, it can be humidity, it can be as I said it can be say for example, earthquake and on all these things can be measure.

(Refer Slide Time: 07:48)

Frequency data

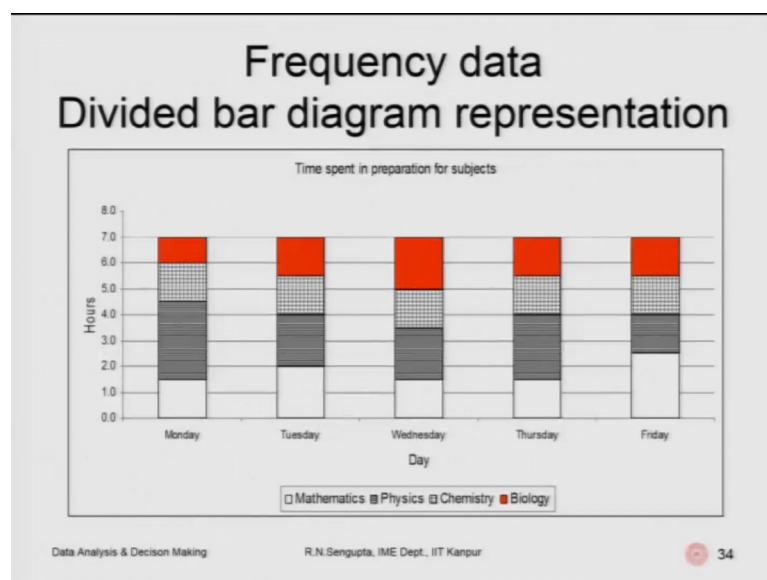
Divided bar diagram representation

Consider we have the time spent in hours by student appearing for the CBSE examination for the preparation of Mathematics, Physics, Chemistry and Biology. We collect the student's preparation pattern for a five day period and want to represent the data thus obtained. In that case we would use the divided bar diagram as illustrated below.

Data Analysis & Decision Making R. N. Sengupta, IIM Dept., IIT Kanpur 33

So, now we will consider the divided bar diagram representation consider we have the time spent in hours by a particular student, who is appearing for the CBSE examination for the pre and he or she is preparing for the following subjects which are Mathematics, Physics, Chemistry and Biology. We basically collect the student's preparation pattern for a five day period and want to represent the data thus are thus obtained.

(Refer Slide Time: 08:15)



So, in that case we would use the divided bar graph as illustrated now. So, in this illustration we have the days which is Monday, Tuesday, Wednesday, Thursday, Friday as

I said is 5 days along the x-axis; along the y-axis, we have basically the hours. And we will consider the person is preparing about 7 hours daily, so that basically is an assumption which we are taking that the person he or she is basically starting each day 7 hours. It can be different, so even if it is different, we can represent it accordingly.

So, if you see Monday, the overall the colored scheme are like this. If it is white, it is mathematics. It is horizontal line it is physics. It is checkered lines box one, it is chemistry; and if it is red, it is biology. So, if you see that in a hours a 7 hours scale, on the Monday the person studies 1 hour for biology, in on Tuesday he or she studies 1 and an half hours, on Wednesday he or she basically studies 2 hours, and on Thursday, and Friday he or she studies one and an half hours each for Biology.

Now, if you consider accordingly for mathematics, it is one and a half. So, I am going from Monday to Friday is one and a half hours, 2 hours, one and an half hours, one and an half hours, and two and an half hours. So, the total sum for each day is basically 7 hours. So obviously, if it is not 7 hours, it can be done on a proportional scale (Refer Time: 09:38).

(Refer Slide Time: 09:41)

**Frequency data
Stacked column diagram
representation**

The method of depicting the data is almost similar to the divided bar diagram representation, but here we represent the percentage wise figures for the variables for each data point. Consider we are finding the consumption in rupees for the four main categories of food of a family in the months of January to June. Remembering that the total amount spent for each month can be different, we depict the percentage wise consumption in food for the four categories.

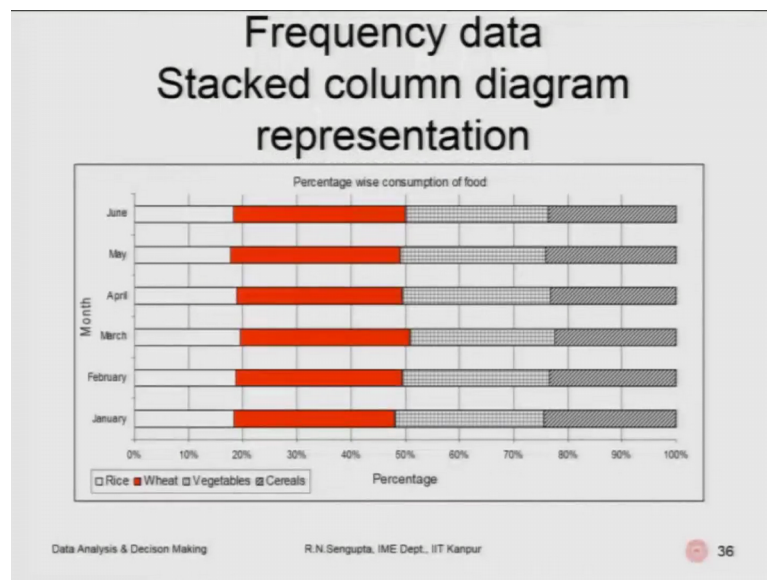
Data Analysis & Decision Making R. N. Sengupta, IIM Dept., IIT Kanpur 35

Now, we consider the stacked column diagram representation. The method of depicting the data is almost similar to the divided bar diagram representation, but here we represent the percentage wise figures for the variables of each data. So, if you remember

just few minutes back or a few seconds back, I mentioned that if the person is studying more or less than 7 hours how it can be represented, so this is the way.

So, consider we are finding the consumption of in rupees for the four main categories of food of a family in the month of January to June January, February, March so and so forth till June. Remember that the total amount spend for each month can be different. So, obviously in January, I may be willing to spend more on say for example, on my total consumption; in February maybe less; March it may be much more and so on and so forth. So, remember that the total amount spent for each month can be different, we depict the percentage wise consumption in food for the four categories accordingly.

(Refer Slide Time: 10:33)



So, this is basically drawn horizontally. So, we have the percentage along which is from 0 to 100 along the x-axis, and for the month which is given as January to June along the y-axis. And the color schemes are accordingly rice is white, wheat is red, checkered one, box one is vegetables, cereals is slanted hashed ones. So, if you see the percentage wise consumption, in rice is almost equal to 20 percent in the month of March; and while for the month of May it is the least. And similarly, if we consider the consumption cereals, it is basically the least in the month of March, and maximum almost maximum for the month of January and May.

(Refer Slide Time: 11:20)

Frequency data

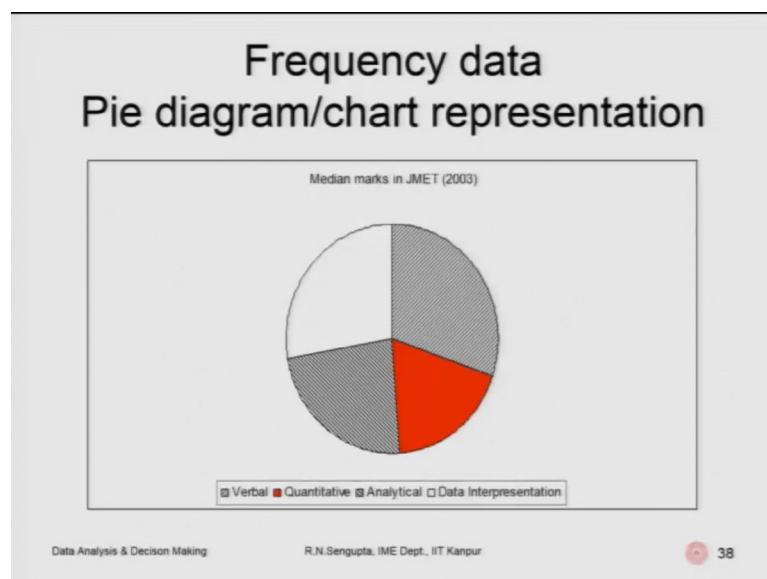
Pie diagram/chart representation

When the values of a variable are given for a number of categories, as in spatial series, we may be interested in a comparison of the categories or series or the contribution of each category to the total. Here the proportions or percentages of various categories, rather than the absolute values for the categories, will be the principal subject of study

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 37

Now, we will consider the frequency diagram representation in the pie chart. So, when the values of a variable are given for a number of categories, as in a spatial series, we may be interested in a comparison in trying to find a comparison of the categories of series or the contribution of each category to the total. Here the proportions of percentage of various categories, rather than absolute values of the categories that are depicted and they would basically be the principal subject of study.

(Refer Slide Time: 11:47)



So, consider this is the JMET examination which was the examination which was held by all the IIT's in order to consider the students for the MBA program. So, now, it is basically replaced by CAT. So, this is the frequency data representation on a pie diagram chart representation. For the median marks in JMET in 2003, so there were 4 categories based on which the JMET was taken. So, it was verbal which is slanted from top right to the bottom left, quantitative is red in color, analytic is basically again slanted from left or top to right bottom, and the data representation is done basically using the white one.

So, if you consider that, the data representation is definitely more than one-fourth of the total pie which is 360 degrees. So, each sector or the set is basically 90 degrees which is more than 90 degrees for the data interpretation part.

(Refer Slide Time: 12:43)

Frequency data
Textual representation

In textual representation of data we depict the information through text.

Consider for the year 2004-2005 we know the number of post graduate students who have registered in different engineering course at IIT Kanpur. The figures are 83 in Aerospace, 88 in Chemical, 139 in Civil, 222 in Electrical, 176 in Mechanical, 115 in Computer Science. Given this data we may be required to utilize this information to answer some queries.

Data Analysis & Decision Making R.N. Sengupta, IIM Dept., IIT Kanpur 39

Is the textual representation, in the textual representation of the data we depict the information through text. So, let me read it you will understand. Consider for the year 2004-2005 we know that the number of postgraduate students who have registered in different engineering courses at IIT Kanpur are as follows. The figures are 83 in Aerospace, 88 in Chemical, 139 in Civil, 222 in Electrical, 176 in Mechanical, 115 in Computer Science and so on and so forth. Given this data we may be required to utilize the information to answer some of the queries, which maybe rise.

Say for example, we may need to find out what is the total number of female students who are there in the different programs. We may be interested to find out that what is the

total percentage in which CAT in which department the percentage of students with the highest along the total group and so or what is the total highest GATE marks for the students who have entered different programs. So, we can utilize this accordingly.

(Refer Slide Time: 13:40)

**Frequency data
Stem leaf representation**

The stem and leaf representation is a quick way of looking at the data set. It contains the information of a histogram but avoids the loss of information in a histogram that results from aggregating the data into intervals. The stem and leaf display is based on the tallying principle but also uses the decimal base of our number system. In the stem and leaf representation, the stem is the number without its rightmost digit (the leaf). The stem is written to the left of a vertical line separating the stem from the leaf. Suppose we have the numbers 105, 106, 107, 109, 100, 108. Then if we use the stem and leaf representation we would depict the numbers as 10 | 567908

Data Analysis & Decision Making R. N. Sengupta, IIM Dept., IT Kanpur 40

Now, I will consider the stem leaf representation. The stem and leaf representation is a quick way of looking at the data set. It contains the information's of a histogram, but avoids the loss of information in a histogram that results from the aggregating the data into intervals. The stem and leaf display is based on the tallying principle, but also uses the decimal base of a number system. In this stem and leaf representation, the stem is the number without its rightmost digit, so which is the basically we will subsume under the leaf.

The stem it is written to the left of a vertical line separating the stem from the leaf. Suppose, we have the numbers as 105, 106, 107, 109, 100, 108, then if we use the stem leaf diagram, it will be 10 which is the stem the slash which will now denote the leaf. So, the leafs are 5, so it would be 105, next one is 6 it will be 106, next one is 7 it will be 107 so on and so forth till the last one which is 108.

(Refer Slide Time: 14:38)

Frequency data Box plot representation

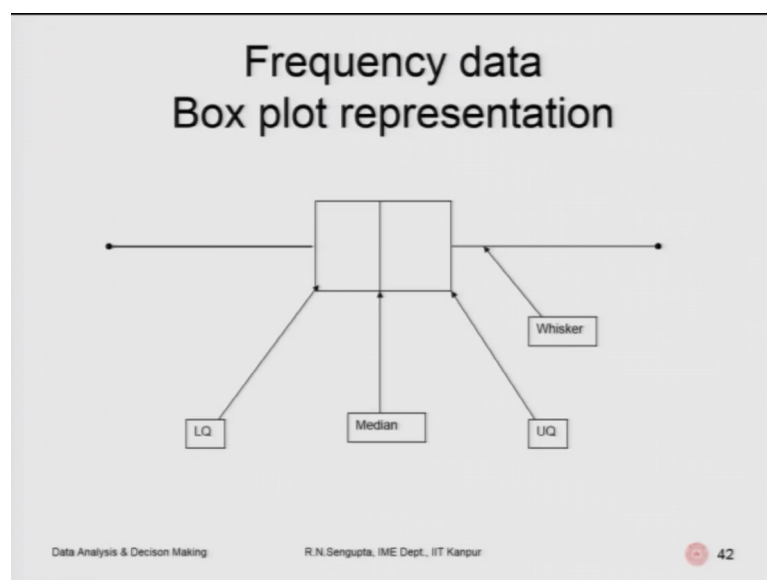
The box plot is also called the box whisker plot. A box plot is a set of five summary measures of distribution of the data which are

- median
- lower quartile
- upper quartile
- smallest observation
- largest observation.

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 41

So, next frequency data representation is the box plot representation. The box plot is to also call the box whiskers plot. A box plot is a set of five summary measures of the distribution that or the data which are basically the information which you will be studying later on; which is the median, the lower quartile, the upper quartile, the smallest in observation the largest observation and so on and so forth.

(Refer Slide Time: 15:03)



So, if we see the diagram, it will be it very rudimentary one. I have trying to drawn the box would basically depend and the median value would be the value at which what

percentage of the probability you are covering on the left, and the right we will come to that later on. The lower quartile and the upper quartile will be given by the left most in the right most, and the whisker would basically extend on to the right on the left depending on the maximum value in the minimum value.

(Refer Slide Time: 15:34)

**Frequency data
Box plot representation**

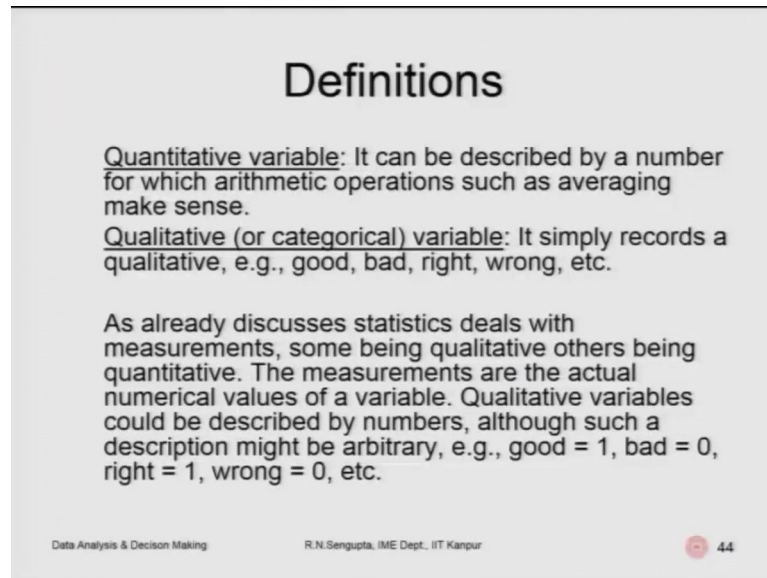
Here:

- $UQ - LQ =$ Inter quartile range (IQR)
- $X =$ Smallest observation within $1.5(IQR)$ of LQ
- $Y =$ Largest observation within $1.5(IQR)$ of UQ

Data Analysis & Decision Making R.N. Sengupta, IIM Dept., IIT Kanpur 43

So, just methodology of our representation, the box plot representations are upper quartile may lower quartile will give you the inter quartile range. X would be the smallest observation which technically can be this is not a sacrosanct number it will be 1.5 into the inter quartile range simply for the largest value it would be 1.5 in the inter quartile range. So, the values of 1.5, 1.5 are the case being utilized. When we think it is a symmetric distribution, obviously if it is a non symmetric distribution, the values would change. Values means 1.5 would be different for both the left quartile left part and the right part.

(Refer Slide Time: 16:06)



Definitions

Quantitative variable: It can be described by a number for which arithmetic operations such as averaging make sense.

Qualitative (or categorical) variable: It simply records a qualitative, e.g., good, bad, right, wrong, etc.

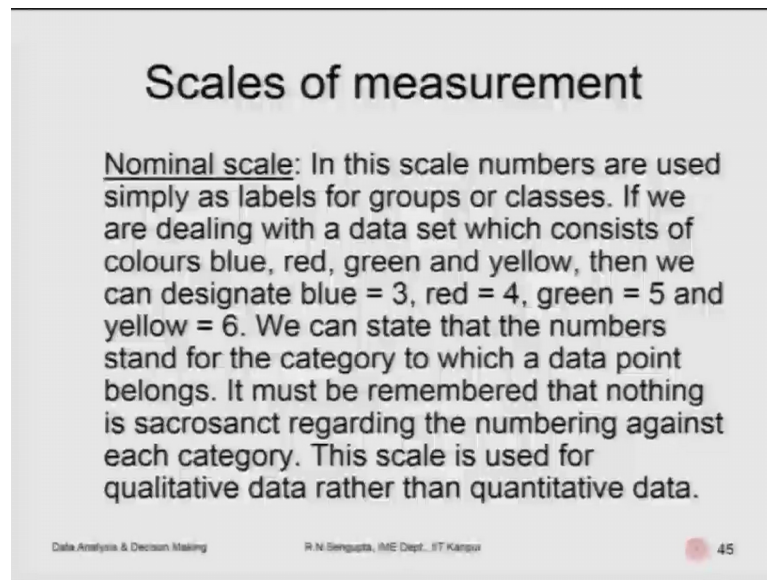
As already discussed statistics deals with measurements, some being qualitative others being quantitative. The measurements are the actual numerical values of a variable. Qualitative variables could be described by numbers, although such a description might be arbitrary, e.g., good = 1, bad = 0, right = 1, wrong = 0, etc.

Data Analysis & Decision Making R.N.Sengupta, IIM Dept., IIT Kanpur 44

So, we will consider a few definitions. So, quantitative variables, it can be predicted by a number on which the arithmetic operations such as the averaging would make sense. While qualitative variables would be the ones in simply some attributes. You see, it simply records a qualitative characteristics like good, bad, right, wrong, color wise, number wise whatever it is.

As already discussed and the statistics deals with measurements some being qualitative others being quantitative. The measurements are the actual numeric values of a variable. Qualitative variables could be described by numbers, although such a description may be arbitrary, say for example, if we denote good as 1 we may be tempted to give bad as 0, it can be minus 2 also, minus 3 also, we do not know, but is basically a and a some judgment would be you utilized, but it is obviously, not always objective some such the subjective would also be there. So, right can be would 1 wrong can be 0, or if wrong is 0 right can be plus 3 also. So, depending on how we have been able to set up the problem we will give numbers accordingly to the attributes.

(Refer Slide Time: 17:16)



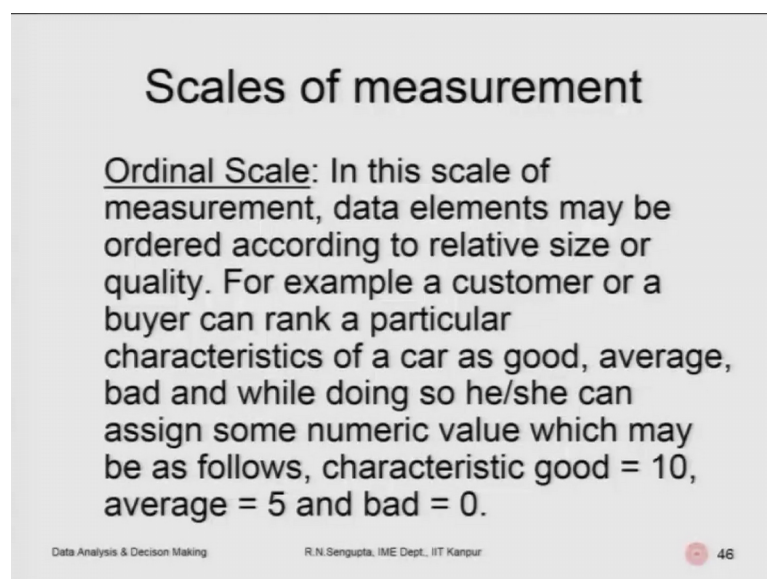
Scales of measurement

Nominal scale: In this scale numbers are used simply as labels for groups or classes. If we are dealing with a data set which consists of colours blue, red, green and yellow, then we can designate blue = 3, red = 4, green = 5 and yellow = 6. We can state that the numbers stand for the category to which a data point belongs. It must be remembered that nothing is sacrosanct regarding the numbering against each category. This scale is used for qualitative data rather than quantitative data.

Data Analysis & Decision Making R. N. Sengupta, IIM Dept., IIT Kanpur 45

Next is the Nominal scales or the scales of measurement. In this scale numbers are utilized simply as labels for groups or classes. If we are dealing with a data set which consists of colors blue, red, green and yellow, then we made give numbers or designate numbers to the colors as blue as 3, red as 4, green as 5 and yellow as 6. We can state that the number stand for the category to which the data point belongs. So, it may be remembered that nothing is sacrosanct the concept of the numbers which we give to the colors regarding the numbering against each category or the characteristics. This scale is used for qualitative data rather than quantitative data.

(Refer Slide Time: 17:53)



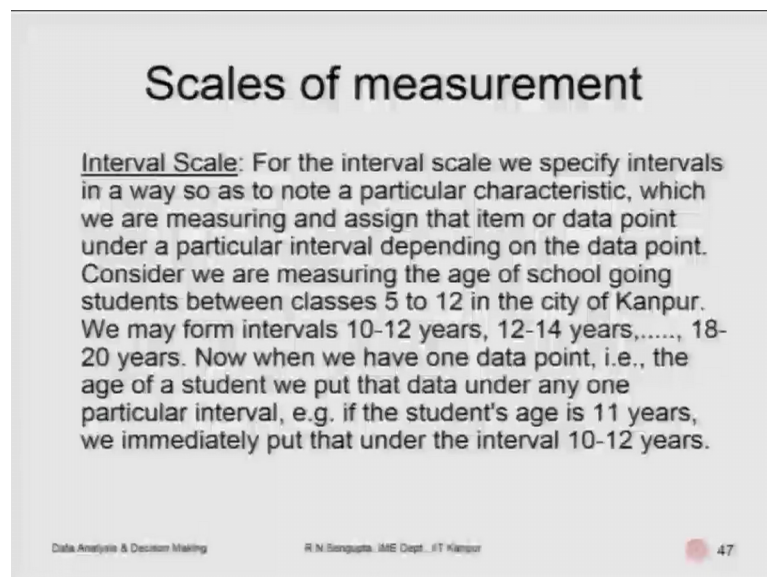
Scales of measurement

Ordinal Scale: In this scale of measurement, data elements may be ordered according to relative size or quality. For example a customer or a buyer can rank a particular characteristics of a car as good, average, bad and while doing so he/she can assign some numeric value which may be as follows, characteristic good = 10, average = 5 and bad = 0.

Data Analysis & Decision Making R. N. Sengupta, IIM Dept., IIT Kanpur 46

Ordinal scale, see the second concept of the scales of measurements. In this scale of measurement, data may be ordered according to relative size or quality. For example, a customer or a buyer can rank a particular characteristics of car as good, average, bad not so good and while doing so he or she can may assign some numeric values which can be as follows like characteristics is good give a score of 10, if it is average give a score of 5, it is bad give a score of 0. So, this 10, 5, 0 or whatever the number is also subjective, they may not be any actual theoretical objectivity in the ranking system.

(Refer Slide Time: 18:34)



Scales of measurement

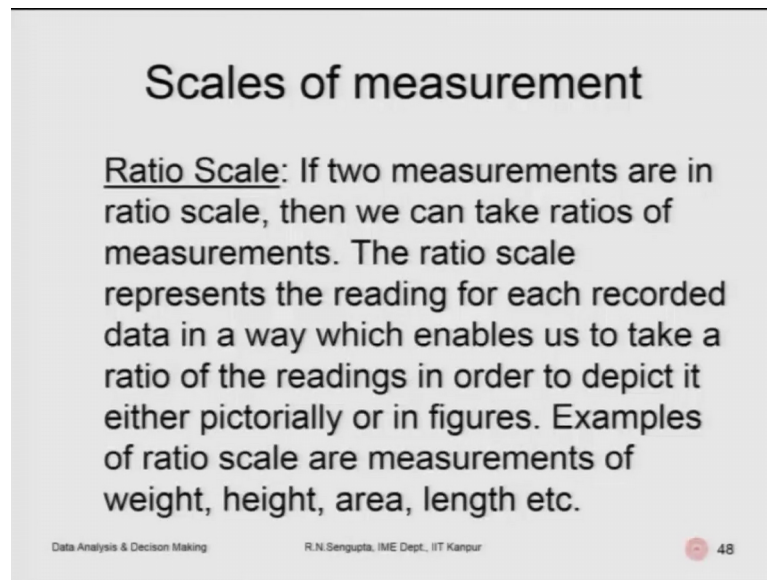
Interval Scale: For the interval scale we specify intervals in a way so as to note a particular characteristic, which we are measuring and assign that item or data point under a particular interval depending on the data point. Consider we are measuring the age of school going students between classes 5 to 12 in the city of Kanpur. We may form intervals 10-12 years, 12-14 years,....., 18-20 years. Now when we have one data point, i.e., the age of a student we put that data under any one particular interval, e.g. if the student's age is 11 years, we immediately put that under the interval 10-12 years.

Data Analysis & Decision Making R. N. Singhania, JME Dept., IIT Kanpur 47

Scales of measurements is basically we will consider the interval scales. For this interval scales, we specify intervals in which as to note a particular characteristics, which we are measuring and assign that item or data point under a particular interval depending on the data point. Consider by measuring the age of the school going students between the class 5 to 12 in the city of Kanpur.

We may form intervals like 10 to 12, 12 to 14 and so on and so forth till the group of 18 to 20. Now when we have one data point that is the age of the student we put that data or that student under any one particular interval, example in the student's age is 11 we immediately put that student under the interval 10 to 12.

(Refer Slide Time: 19:17)



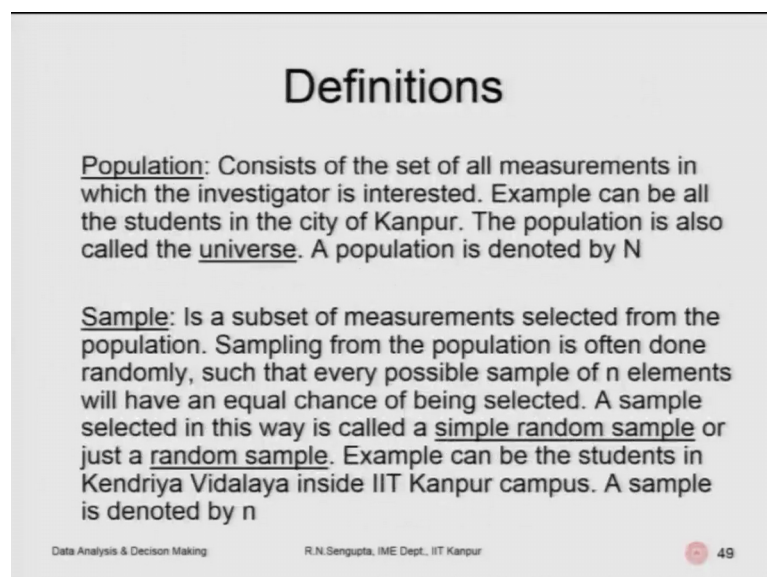
Scales of measurement

Ratio Scale: If two measurements are in ratio scale, then we can take ratios of measurements. The ratio scale represents the reading for each recorded data in a way which enables us to take a ratio of the readings in order to depict it either pictorially or in figures. Examples of ratio scale are measurements of weight, height, area, length etc.

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 48

Now, in the next concept of scale of measurement will be the ratio scale. If two measurements are in the ratio scale, then we can take the ratios of the measurements. The ratio scale represents the reading for each recorded data in a way which enables us to take a ratio of the reading in order to depict it either pictorially or in figures. Examples or ratio scales are measurements of weight, height, area, length, volume and so on and so forth.

(Refer Slide Time: 19:44)



Definitions

Population: Consists of the set of all measurements in which the investigator is interested. Example can be all the students in the city of Kanpur. The population is also called the universe. A population is denoted by N

Sample: Is a subset of measurements selected from the population. Sampling from the population is often done randomly, such that every possible sample of n elements will have an equal chance of being selected. A sample selected in this way is called a simple random sample or just a random sample. Example can be the students in Kendriya Vidyalaya inside IIT Kanpur campus. A sample is denoted by n

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 49

Now, we will consider something of our more conceptual sense. So, we will consider the concept of population and then consider the concept of sample also. So, these what a very brief way maybe I have gone a little bit fast, but is this different ways of depicting different type of data in the frequency type and the non frequency type. Now, whenever we are studying statistics the basically it means that we have a huge population or view set or our universal set and based on this.

We will try to basically understand what the what is the characteristic of the population and the characters of population may not be known to us always with I would not use the word certainty, but is not known to us in the sense, that we may not have all the actual readings from the population based which we can draw some meaningful conclusions.

So, obviously, to mean then what should we do so obviously, we have to pick up a chunk from the population which will denote as a sample, so this is what we are going to basically analyze. So, populations consists on the set of all measurements in which the investigator is interested. For example, can be all the students in the city of Kanpur. The population is also called the universe. A population is denoted by the symbol N . While a sample is a subset of the measurements selected from the population N , Sampling from the population is often done randomly, such that every possible sample of n elements, so this small n is basically the sample size which I am picking up.

So, every possible sample of size n will have an equal chance of being selected. So, if we say for example, if you have a box, and which are marked from 1 to 100 consider that is the population. So, consider that chits are 1 to 100 each chit is only once. So, basically we have the population as 100 such observation technically that is a thought out experiment. Now you want to basically pick up a sample and sample size of 10. Now why it is saying that each has a possible being or n elements and will have the equal chance of being selected like this, and let us basically go into detail see for example, if I pick up the observation, which is mark the chit 1.

So obviously, if we pick up the chit one, the first time the probability is $1/100$. Now if he basically removed the chit, so hence the total population technically would, now be not 100 it would be 99, and hence the corresponding probability of picking up any other chit, basically, now becomes $1/99$. So obviously, if we say that what is the probability of picking up number 2, it is $1/99$, picking up a number chit number 55 is $1/99$, but

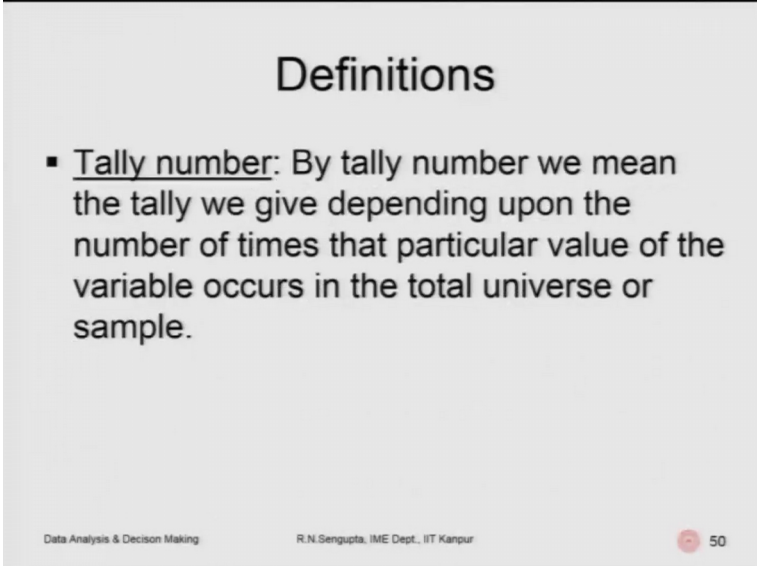
also see on the other hand if I say what is the probability of picking up the chit 1, it is basically now 0, because that has been already removed.

But now, if we change the experiment in such a way that if we pick up the chit1, noting down and again put it back in the box. So obviously, the population remains same technically it means the population is infinite. So, as that the corresponding probability of picking up the any chit is still remains 1 by 100 says that if I again pick up, and in the number 1 comes the probability means to 1 by 100.

Now this is basically known as simple random sampling with replacement and without represent obviously, it would mean that the corresponding probability would start decreasing, but in the if actually if you have a population, which is infinite then, if you pick up chits with replacement and without replacement, the underlying distribution should not change that is the basic main essence.

So, sample selected in this way is called the simple random sampling sample or just a random sample example, can be the students in Kendria Vidyalaya inside IIT Kanpur campus, a sample would basically we denoted by small n.

(Refer Slide Time: 23:30)



Definitions

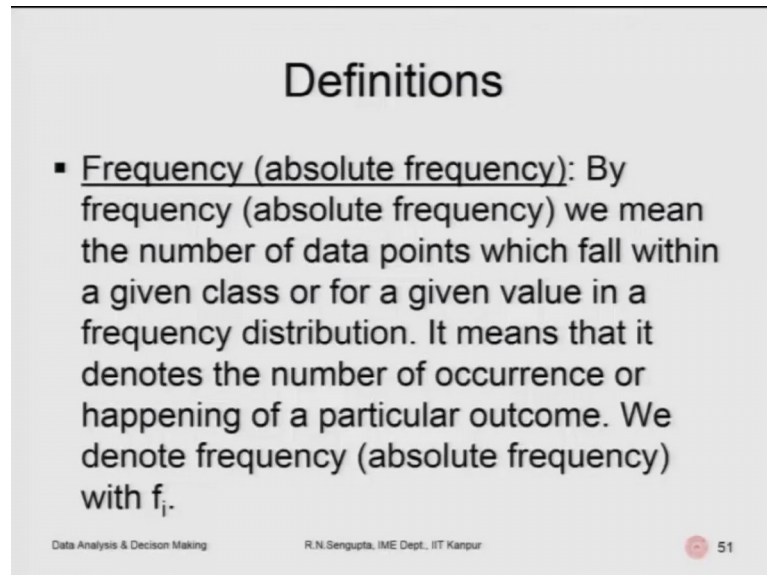
- **Tally number:** By tally number we mean the tally we give depending upon the number of times that particular value of the variable occurs in the total universe or sample.

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 50

Now, what is tally numbers? By tally numbers we mean the tally which we had we give depending upon the number of times at a particular value or the variable occurs in the total universe or sample. And this tally numbers basically we will give you in depending

on the on the value will basically count the number of tally numbers, and basically say what is the frequency of the occurrence of a particular value.

(Refer Slide Time: 23:54)



The slide is titled "Definitions" and contains a single bullet point defining absolute frequency. At the bottom, it includes the text "Data Analysis & Decision Making", "R.N.Sengupta, IIM Dept., IIT Kanpur", and a red circle with the number "51".

Definitions

- Frequency (absolute frequency): By frequency (absolute frequency) we mean the number of data points which fall within a given class or for a given value in a frequency distribution. It means that it denotes the number of occurrence or happening of a particular outcome. We denote frequency (absolute frequency) with f_i .

Data Analysis & Decision Making R.N.Sengupta, IIM Dept., IIT Kanpur 51

Now, we will also consider these are the definitions also consider the concept of frequencies which is absolute frequency by frequency, which is absolute frequency we mean the number of data points which fall within a given class or for a given value in a frequency distribution. It means that it denotes the number of occurrences or happening of a particular outcome. We denote the frequency which is the absolute frequency by the symbol of small f , suffix i where i basically denotes the number reading.

So, obviously, if there are three different type of numbers, and if the chit has basically 100, such numbers occurring 100 means 100 number such papers which are there. But the number occurring on numbers which are occurring is only 1 or 2 or 3. So, consider number 2 occurs 50 times, number 1 occurs 25 times, and number 3 occurs 25 times.

So, if basically I am saying the frequency of number 1 occurring, it is basically 25, frequency of 2 occurring is basically 50. So, that is why it will be f suffix 1 which is for number 1 is 25, f suffix 2 which is num for number 2 is 50, and f suffix 3 which is of for number 3 is 25.

(Refer Slide Time: 25:09)

Definitions

Cumulative frequency: The cumulative frequency corresponding to the upper boundary of any class interval or value in a frequency distribution is the total absolute frequency of all values less (greater) than that boundary for the class or value. We denote cumulative frequency less (greater) than type by

$F = \sum_{i \leq n} f_i$ $F = \sum_{i \geq n} f_i$

Data Analysis & Decision Making R.N.Sengupta, IIM Dept., IIT Kanpur 52

So, we will consider the concept of cumulative frequency. So, the cumulative frequency corresponding to the upper boundary of any class interval or value, in a frequency distribution and it is basically the total absolute frequency of the all the values either of the less than type of the greater than type, what is less than type and greater than type. I am going to come to that within few minutes.

Then the boundary for the class of the value would basically be given by the cumulative frequency. So, we denote the cumulative frequency of the less than type and the greater than type by the corresponding formula which is, so this capital F which basically let me, so the capital F basically denotes the frequency of the less than type if it is this, so obviously that will be all the sum of all the frequencies for the values of y less than equal to n.

So, this is the less than type value and if I am denoting in the concept of the greater than type, so let me denote in the colors, so the greater than type would be denoted by a frequency of capital F greater than n which means for all the values are i which is greater than and less than n, and it will basically mean the summation of all the frequencies for all the values of y which are greater than n.

So, that means, if I am saying that the frequency of the number of number of ones are less than 1, it will be 25 if I am saying the frequency of numbers of less than 2, it will be 25 plus 50, but when I am saying in the frequency of the numbers of greater than 3, so

obviously, it will be 0. So, if I am saying that is the frequency of the less than 2, it obviously it would be 2. It will mean the only the numbers of 1 only which is 25.

(Refer Slide Time: 27:01)

Example

Consider we have the following data related to the size in numbers of thirty families in the city of Jaipur.

2, 6, 3, 4, 4, 5, 3, 6, 4, 4, 5, 3, 2, 3, 6, 5, 4, 4, 4, 3, 2, 4, 5, 6, 7, 4, 4, 5, 3, 3.

Now the question is how do we represent the data using tally numbers, frequencies, cumulative frequencies?

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 53

So, consider the example. We have the following data related to the size in the numbers of 30 families in the city of Jaipur. So, they are given as numbers 2, 6, 3, 4, 4 and so on and so forth till the number 3 and 3. So, now, there are 30 families. Now, the question is how do we represent the data using the tally numbers with the frequencies of the cumulative frequencies? So this is how we do it.

(Refer Slide Time: 27:23)

Example

# of members	Tally #	f_i	$F = \sum_{i \leq n} f_i$	$F = \sum_{i \geq n} f_i$
2		3	3	30
3		7	10	27
4		10	20	20
5		5	25	10
6		4	29	5
7		1	30	1

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 54

So, on the left most column, we write the number of families members which are there in the family. So, it can be minimum is 2 and the maximum is 7. So, if the tally numbers are given, so it means 3 which is the frequency for 2 number of members being a family is 3. If you are trying to find out the tally number for number of family members which is exactly 3, it is 7, if family members number of exactly 4 is 10, for 5 is 5, 6 is 4, and 7 size being 1.

If I am trying to find out the frequency of less than time, it will basically be number of families which are less than equal to 2 is 2 is 3; family members which is less than equal to 3 it will be 3 plus 7 which is 10; family size of less than 4, less than equal to 4 would be, families having 2, families having 3, families having 4, which is 20 so on and so forth. So, till the last one which I want to find out is the family size I having members less than 7, less than equal to 7 is 30. If I go in other way around family size of greater than that value, if it is greater than that value of n , so obviously it will be just with the reverse starting from 30 to 1.

So, with this, I will end the second lecture and continuing the discussion more about probability more of frequency relative frequency chance and so on and so forth such that we will slowly go into the concept of probability. So, I with this I will end the second class and hope that you will keep yourself abreast with the discussion. So, we maybe will going a little bit fast, but try to basically cover the initial topics in such a way that we are we have quite a lot of time trying to basically cover the concepts of multivariate statistics and so on and so forth. Have a nice day.

Thank you very much.