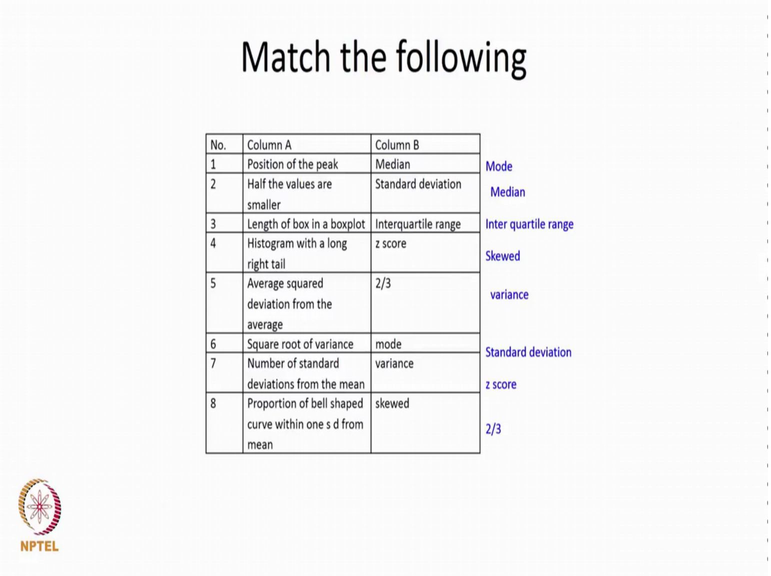
Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 08
Exercises, Association between categorical variables

In this lecture we begin with the discussion on the topic that we saw earlier which is how to describe numerical data and then we will go on to discuss measure of Association among Categorical Variables. So, we first start with describing the numerical data under discussion.

(Refer Slide Time: 00:39)



Match the following

No.	Column A	Column B	
1	Position of the peak	Median	Mode
2	Half the values are smaller	Standard deviation	Median
3	Length of box in a boxplot	Interquartile range	Inter quartile range
4	Histogram with a long right tail	z score	Skewed
5	Average squared deviation from the average	2/3	variance
6	Square root of variance	mode	Standard deviation
7	Number of standard deviations from the mean	variance	z score
8	Proportion of bell shaped curve within one s d from mean	skewed	2/3

So, let us start with the small match the following exercise to understand what happens. So, in this so, we try to match the things in column A versus things in column B. So, there are 8 items that are given. So, position of the peak is the first thing in column A and if we look at the options we realize position of the peak indicates that value which has the largest frequency and therefore, it has to be the mode which we find here under item 6. So, position of the peak will be the mode.

The second one we see half the values are smaller. So, if we look at the alternatives we realize that among these alternatives when we say half the values are smaller we are looking at the middle value and assuming that these values are sorted we are looking at the middle value, so that half the values are smaller. So, the moment we look at middle values it can only be a measure of central tendency. So, a measure of central tendency that we have in this list in column B or median and mode we have already used the mode, but we also know that the median is a measure which is in the middle. So, 50 percent of the numbers or 50 percent of the data are lesser than the median and 50 percent are more than the median. So, half the values are smaller the correct answer is the median.

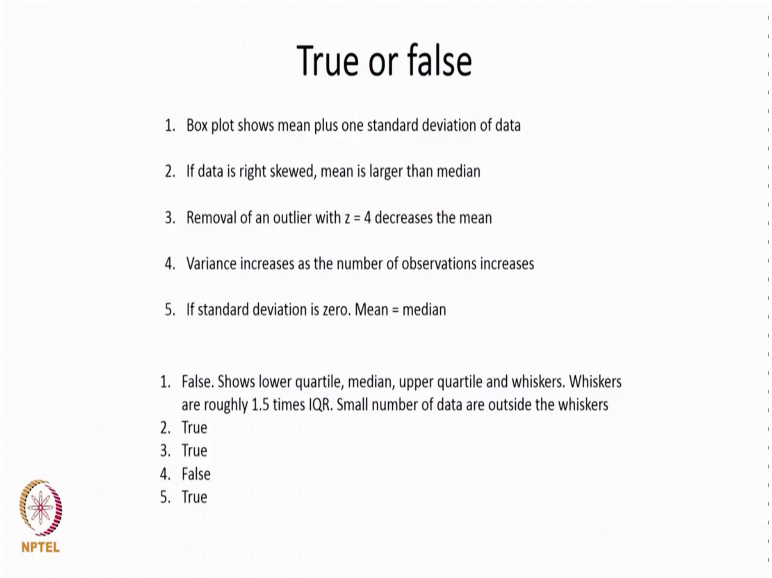
Length of the box in a box plot; we observe that it is the interquartile range we saw when we discussed in the last lecture that we can find the median in the box plot and then we can find the lower and upper quartile and therefore, the length of the box plot is the interquartile range.

Histogram with the long right tail; example, the salary data so, histogram with the long right tail among all these options is skewed. So, either when the right tail is longer or the left tail is longer we say it is skewed. So, in this case the correct answer is skewed. Average squared deviation from the average. So, given a set of numbers we calculate the average or the arithmetic mean and then we find out the difference between every number and the arithmetic mean and square it and that in this answer is the variance. Because we also find the average of the sum of the squared deviations about the mean and therefore, the correct answer is the variance.

The next one is the square root of the variance. This is a very simple square root of the variance we have already seen is the standard deviation and therefore, that is the answer. So, number of standard deviations from the mean which is given by a z score we have not looked at the z score yet I am just introducing this idea in a normal distribution which we saw the bell shaped curve there is a z score which is called x minus mu by sigma where mu is the mean and sigma is the standard deviation. So, number of standard deviations from the mean is called the z score which we introduce now using this.

And, proportion of the bell shaped curve within one standard deviation from the mean is 2 by 3, we said about 68 percent. So, it is 2 by 3. So, this helps us to understand the basic concepts that we saw in the previous lecture.

(Refer Slide Time: 04:20)



True or false

1. Box plot shows mean plus one standard deviation of data
2. If data is right skewed, mean is larger than median
3. Removal of an outlier with z = 4 decreases the mean
4. Variance increases as the number of observations increases
5. If standard deviation is zero. Mean = median

1. False. Shows lower quartile, median, upper quartile and whiskers. Whiskers are roughly 1.5 times IQR. Small number of data are outside the whiskers
2. True
3. True
4. False
5. True

Now, let us look at some true or false questions to see whether we have understood things well. So, box plot shows the mean plus one standard deviation of data. The answer is also given in the same slide, but we will look at the answer after a discussion.

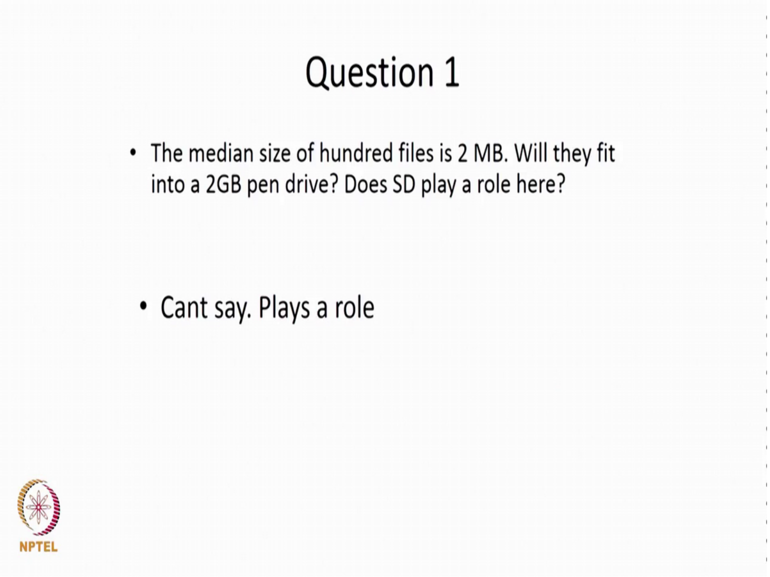
So, if we go and understand what the box plot is the box plot starts with the median and does not start with the mean and therefore, the answer has to be false. We also saw the range in the box plot which is the inter quartile range and the box plot does not talk about the mean therefore, it cannot discuss standard deviation and therefore, the answer is false. It only shows the lower quartile median upper quartile and whiskers. So, whiskers are roughly 1.5 times the IQR which means all the data that is outside the IQR or IQR is called Interquartile Range and those are the whiskers and some data are outside the interquartile range, but a very small number of data are actually even outside the whiskers.

If data is right skewed the mean is larger than the median, the answer is true. We have shown that in an earlier slide in an earlier lecture. Removal of an outlier with z equal to 4 decreases the mean; actually we have not yet discussed z equal to 4 in great detail, but I did make a mention that z in the previous slide that z comes from the normal distribution. So, z equal to 4 in a bell shaped curve is a point which is pretty much to the right which is much higher than the mean and therefore, if we remove a number which is much higher than the mean it is quite likely that the new computed mean reduces and therefore, the answer is true.

Variance increases as the number of observations increases one can always give a counter example. Suppose, I have 5 numbers and I find out the variance now I include a sixth number which is equal to the arithmetic mean and if I do that the contribution of the sixth number to the variance is 0, but the denominator increases with the addition of a number and the variance can decrease. Therefore with a counter example one can say that this can be false.

If the standard deviation is 0, then mean is equal the median. So, when will the standard deviation be 0? Standard deviation is 0 when all the values are the same even if one value is different then there will be a positive standard deviation. So, standard deviation is 0 implies all the values are the same which is equal to the mean which is also equal to the median and therefore, the answer to this question is true.

(Refer Slide Time: 07:23)



Question 1

- The median size of hundred files is 2 MB. Will they fit into a 2GB pen drive? Does SD play a role here?

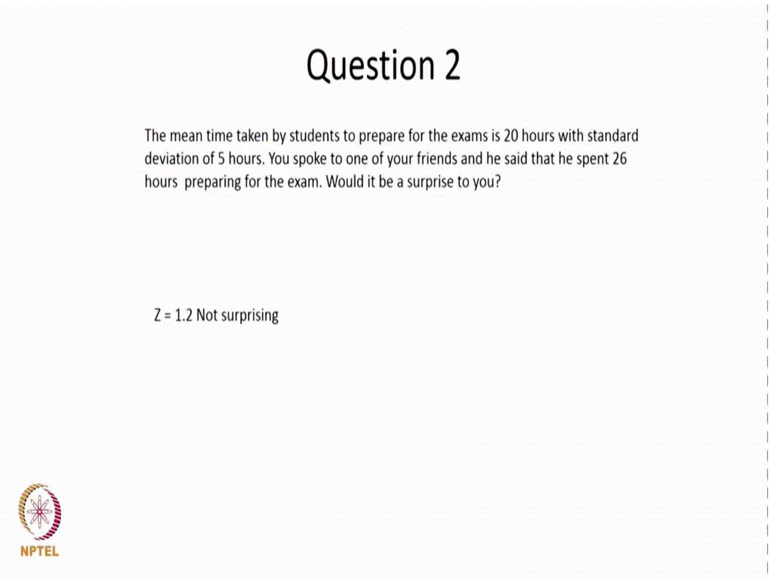
- Cant say. Plays a role

So, now let us look at a few more simple questions to understand this. Now, let us look at a computer and then say that the median size of hundred files is 2 MB. Will they effect into a 2 GB pen drive does standard deviation play a role here?

Now, the answer is also given here we cannot say because median only talks says that 50 percent of these 100 files or 50 files have a size of less than 2 MB, where there could be one which is high and which could run into a 4 GB or whatever it is. So, it does not talk about how large the largest file is therefore, we cannot say that all these 100 files can be put into a 2 GB pen drive and so on.

Does standard deviation play a role here? Yes, standard deviation plays a role here, because if we instead of saying that the median sizes 2 MB if we said that the mean is 2 MB then we know that the total is 200 MB and then we can make a decision about putting it into a 2 GB pendrive. Therefore, the average plays a role the standard deviation also plays a role if there is one file which is larger than 2 GB then the standard deviation of this will also be very high.

(Refer Slide Time: 08:45)



Question 2

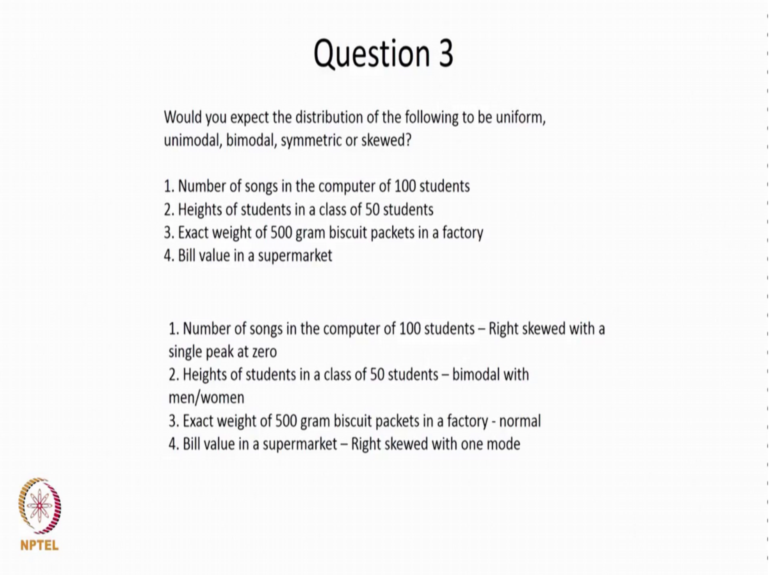
The mean time taken by students to prepare for the exams is 20 hours with standard deviation of 5 hours. You spoke to one of your friends and he said that he spent 26 hours preparing for the exam. Would it be a surprise to you?

Z = 1.2 Not surprising

Let us look at the next question. The mean time taken by students to prepare for the exam is 20 hours with a standard deviation of 5 hours. You spoke to one of your friends and he said that he spent 26 hours preparing for the exam. Would it be a surprise? The answer is maybe not because if we assume that this normal distribution about which we will see in more detail as we move along, but we saw the bell shaped curve and then we concluded that about 68 percent or two-thirds roughly or within one standard deviation on either side.

So, if we look at a large number of people preparing for the exams the mean being 20 and standard deviation being 5 so, you expect about two-thirds of them to spend between 15 and 25 hours preparing for the exam. So, if there is a person who has spent 26 hours it just means that this person is outside of this two-thirds, but within that other one-third and if it is symmetric half of one-third. So, this student can be within the top 20 percent or 18 percent and it may not be very surprising because your friend could be a very studious person who spends more time preparing for the examination

(Refer Slide Time: 10:07)



Question 3

Would you expect the distribution of the following to be uniform, unimodal, bimodal, symmetric or skewed?

1. Number of songs in the computer of 100 students
2. Heights of students in a class of 50 students
3. Exact weight of 500 gram biscuit packets in a factory
4. Bill value in a supermarket

1. Number of songs in the computer of 100 students – Right skewed with a single peak at zero
2. Heights of students in a class of 50 students – bimodal with men/women
3. Exact weight of 500 gram biscuit packets in a factory - normal
4. Bill value in a supermarket – Right skewed with one mode

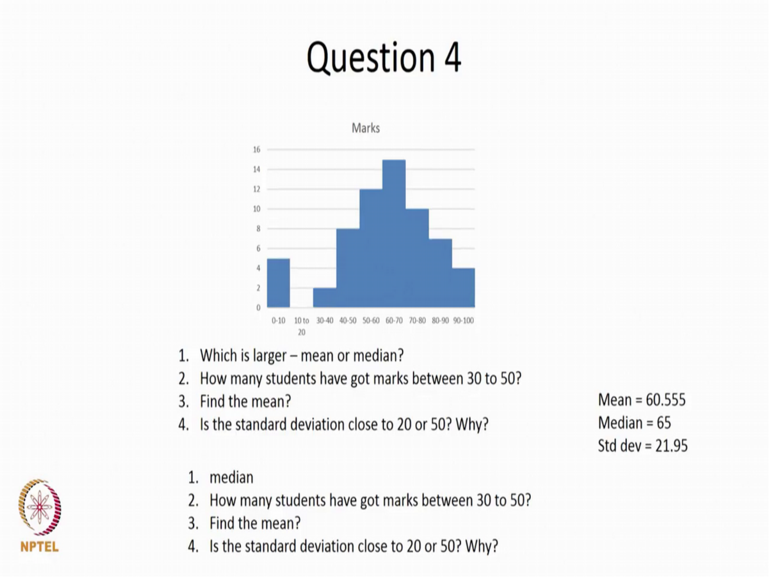
Would you expect the distribution of the following to be uniform, unimodal, bimodal, symmetric or skewed? So, we first have to they have also given the answers below for a ready reckoner. So, uniform distribution means roughly they are all of the same size, unimodal there is a single mode, bimodal there are two modes, symmetric the normal was symmetric about the mean, skewed which is represents the tail.

So, a number of songs in the computers of 100 students the general expectation would be the number of songs in a computer 100 students could be right skewed with a single peak at 0. It is quite likely that more than a good number of this 100 may not have a song and among those who have songs 1 or 2 may have a large number of songs and therefore, there can be a long tail to the right. So, it will be right skewed. So, if we make an assumption that a large number of students would not be having a song in the computer then there will be a single peak at song equal to 0.

Heights of students in a class of 50 students heights of students could be bimodal because have not said how many of them are men and how many of them are women. It is quite likely that the average height of men would be higher than the average height of women so, it could represent a bimodal distribution.

Exact weight of 500 gram biscuit packets we already have seen an example about packing and so on. So, you could expect in this case to be normal with a reasonable peak and a smaller variation, but there will be a variation. Bill value in a supermarket; bill value in a supermarket could be right skewed with one mode. Let us assume this will be this will be quite similar to the number of songs, but then we will not have a peak at 0, we will have a peak at some small range which is there, there can be 1 or 2 small number of customers who would have bought for a large amount of money. Therefore, it could be right skewed with the single mode with small range showing a very high peak and so on.

(Refer Slide Time: 12:23)



Question 4

1. Which is larger – mean or median?
2. How many students have got marks between 30 to 50?
3. Find the mean?
4. Is the standard deviation close to 20 or 50? Why?

Mean = 60.555
Median = 65
Std dev = 21.95

1. median
2. How many students have got marks between 30 to 50?
3. Find the mean?
4. Is the standard deviation close to 20 or 50? Why?

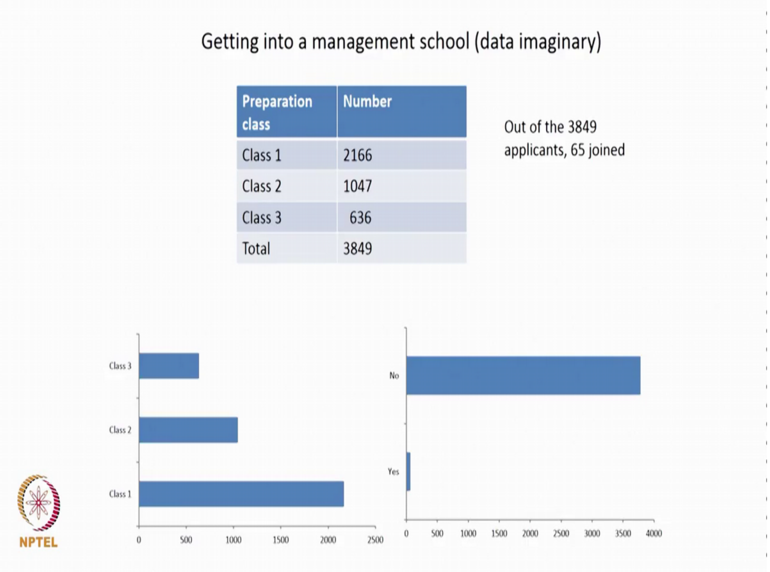
In this question we show a distribution of marks in the form of a histogram which is shown here and the questions are, which is larger the mean or the median? How many students have got marks between 30 and 50? Find the mean and then is the standard deviation close to 20 or 50 and why? So, one is we have also shown the mean median and standard deviation here, but at times by looking at the picture we will be able to say a few things.

In this case the median is 65 the frequency 60 to 70 is the middle frequency in this case in terms of data points and the median is the midpoint of this which is 65, the mean happens to be 60.555. A general look at the data gives us a feeling that it is actually skewed a little bit to the left in a sense there are more points with higher values on the right hand side and therefore, we could have a case where the median is actually higher than the mean is would be somewhere here on the left hand side and the median will be higher than the mean in this case.

How many students have got marks between 30 and 50? So, 30 to 40 there are two students 40 to 50 there are eight students. So, 10 students have got marks between 30 and 50. Find the mean we can actually calculate the mean. So, this is 0 to 10 the frequency is 5, 30 to 40 the frequency is 2 and so on. So, we take each of this range and take the midpoint. So, 0 to 10 is represented by a midpoint of 5 and then the frequency is 5. Here the midpoint is 35, the frequency is 2 and so on and then we can multiply the midpoint with the frequency and then divide it by the frequency which is the number of observations. We will get the mean which happens to be 60.555 in this example.

Is the standard deviation close to 20 or is it close to 50? In this case our answer shows that the standard deviation is 21.95 which is actually closer to 20 then it is to 50. And, one can also try to calculate the standard deviation indirectly. So, direct standard deviation calculation would be sigma f d square by m, where f is the frequency d is the deviation between the mean and the midpoint of this range and we can calculate the standard deviation and if we do. So, we observe that the standard deviation in this case is actually closer to 20 than it is to 50.

(Refer Slide Time: 15:41)



Getting into a management school (data imaginary)

Preparation class	Number
Class 1	2166
Class 2	1047
Class 3	636
Total	3849

Out of the 3849 applicants, 65 joined

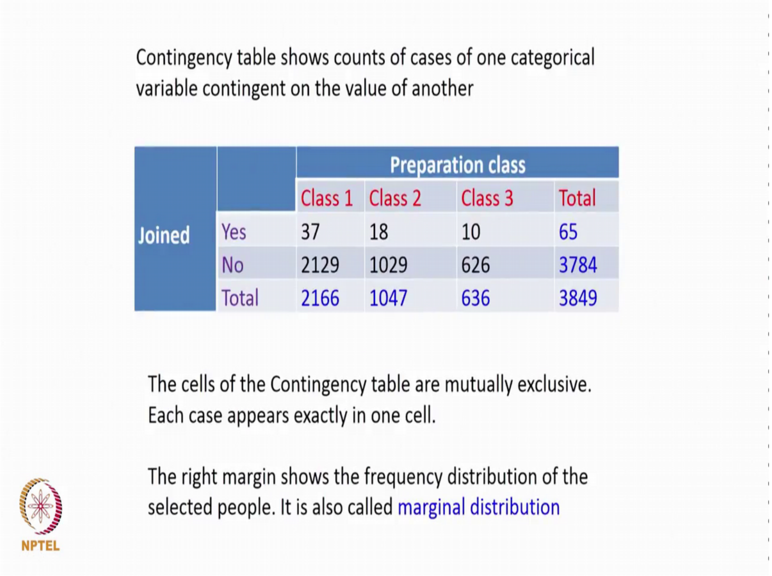
Now, we move to another topic which is association between categorical variables we will start this topic in this lecture and then we will continue this topic in the next lecture.

Now, let us look at some data and try to understand association between categorical variables. Now, we have to go back to look at categorical variables. We have spent so much of time with numerical variables. So, we need to go back to categorical variables and we take an example to understand that. Now, let us assume that we look at students who have gained admission to a management school.

Let us also assume that there were 3849 applications and let us assume that 65 people finally, joined the program. Now, let us also assume that each one of these 3849 students have actually gone to some classes as part of preparation for the admission to the management program and let us say that we consider three classes, class or institute number 1, 2 and 3 which we generic we use a generic expression called class 1, class 2 and class 3.

So, 2166 people went to class 1, 1047 to class 2 and 636 to class 3. So, the bar chart shows that class 1 has 2166 and so on and we also have this case where out of these 3849 know which means people could not or did not join was 3849 less 65 and those who join which is a small bar here which says, yes, are the people who actually joined.

(Refer Slide Time: 17:21)



Contingency table shows counts of cases of one categorical variable contingent on the value of another

		Preparation class			
		Class 1	Class 2	Class 3	Total
Joined	Yes	37	18	10	65
	No	2129	1029	626	3784
	Total	2166	1047	636	3849

The cells of the Contingency table are mutually exclusive. Each case appears exactly in one cell.

The right margin shows the frequency distribution of the selected people. It is also called marginal distribution

Now, let us go back to this data in the form of a table and then we have two values for this joint, yes and no and we have three values or three variables for the preparation classes which are class 1, class 2, class 3. So, the data that we look at are this, the total is 3849, 65 people joined, 3784 either did not or could not join.

Now, this data is further split into this 65 is split into 37 who had gone to class 1, 18 who had gone to class 2 and 10 who had gone to class 3 and then we realized that 2166 in total had gone to class 1, out of which 37 got into the program and 2129 did not get into the program. Similar numbers are 18, 1029 and 10 and 626.

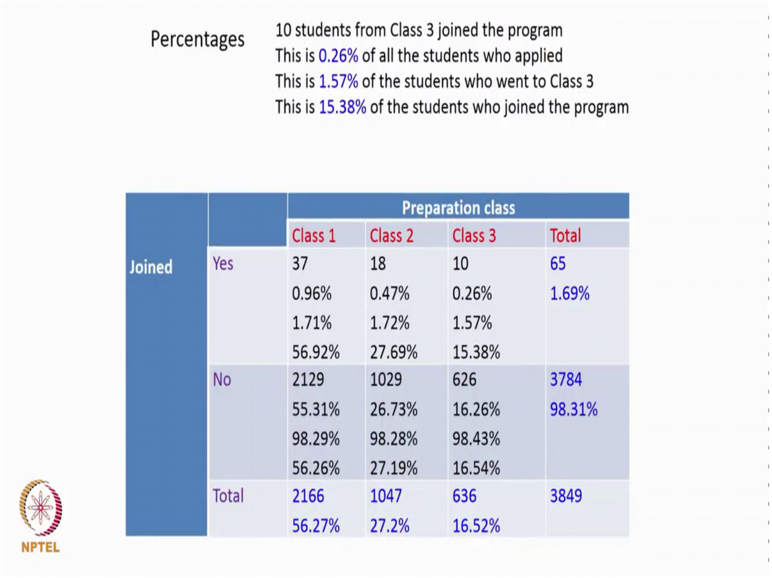
Now, what can we do with this data and what can we understand from this data first the cell is of this is called a contingency table where we try to associate two categorical variables. One categorical variable is joining and not joining and the other variable is the class that they attended prior to joining and not joining.

So, cells are these positions there are 1 2 3 4 5 6 cells in this and this table is called a contingency table. So, the contingency table shows the counts of cases of one categorical variable contingent on the value of another. So, if yes is a categorical variable and contingent on another variable called class 1, which means the number of people who attended class 1 prior to the admission and joined the program the number is 37.

Similarly, we can explain the remaining five numbers. So, cells of this contingency table are mutually exclusive. Each case depends appears exactly in one cell. Now, the total 3849 is the sum of these six numbers and the column sums represent the total in each case which adds up to 3849, the row sum also adds up to 3849.

The right margin shows the frequency distribution of the selected people. It is called marginal distribution 65 out of 3849, 3784 out of 3849 and so on.

(Refer Slide Time: 20:07)



Percentages

10 students from Class 3 joined the program
This is 0.26% of all the students who applied
This is 1.57% of the students who went to Class 3
This is 15.38% of the students who joined the program

		Preparation class			
		Class 1	Class 2	Class 3	Total
Joined	Yes	37 0.96% 1.71% 56.92%	18 0.47% 1.72% 27.69%	10 0.26% 1.57% 15.38%	65 1.69%
	No	2129 55.31% 98.29% 56.26%	1029 26.73% 98.28% 27.19%	626 16.26% 98.43% 16.54%	3784 98.31%
	Total	2166 56.27%	1047 27.2%	636 16.52%	3849

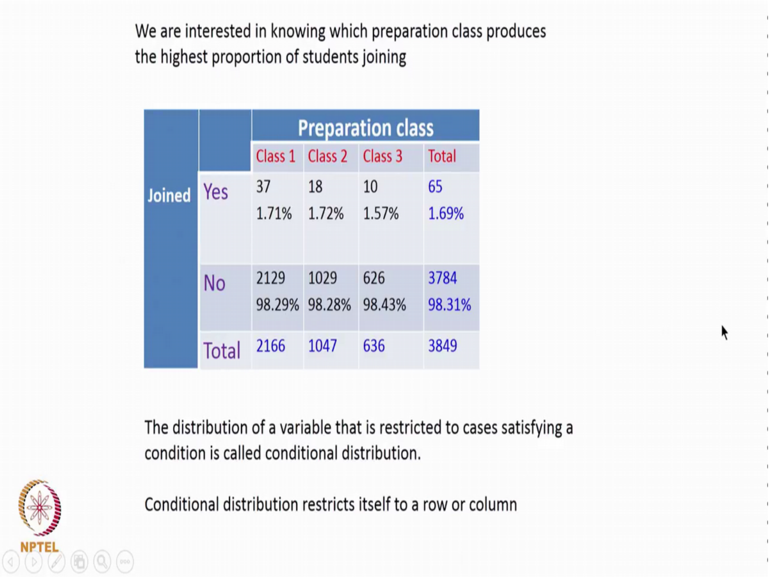
Now, we can represent this thing in the form of percentages. Now, we have shown this table the table has become a little bigger because we have written down all the percentages. Now, let us just explain one out of these six and then we can understand the rest of them.

If we look at this particular block or this particular position, now 10 students were able to come into the program from class 3. So, 10 students from class 3 joined the program. Now, this is 0.26 percent of all the students who applied. So, 0.26 percent of 3849, this is 1.57 percent of those who went to class 3. So, those who went to class 3 is 636. So, 10 by 636 is 1.57 percent. This is 15.38 percent of the students who joined the program.

So, 10 divided by 65 is 15.38. So, we have these three ratios or percentages which are given here for the case joined the program and class 3. The first one the number who went to class 3 and join the program out of all the total, the number who went to class 3 and joining the program out of all those who went to class 3 and the number who went to class 3 out of these people total who joined are also given here. So, typically 10 divided by this total 10 divided by this total and 10 divided by this total.

So, if I look at this for example, the first one will be 2129 divided by 3849 which is 55.31 the next would be 2129 divided by 2166 which is this total which is 98.29 and the third is 2129 divided by 3784 which is 56.26. So, we can compute all these percentages from the table that we actually have.

(Refer Slide Time: 22:23)



We are interested in knowing which preparation class produces the highest proportion of students joining

		Preparation class			
		Class 1	Class 2	Class 3	Total
Joined	Yes	37 1.71%	18 1.72%	10 1.57%	65 1.69%
	No	2129 98.29%	1029 98.28%	626 98.43%	3784 98.31%
	Total	2166	1047	636	3849

The distribution of a variable that is restricted to cases satisfying a condition is called conditional distribution.

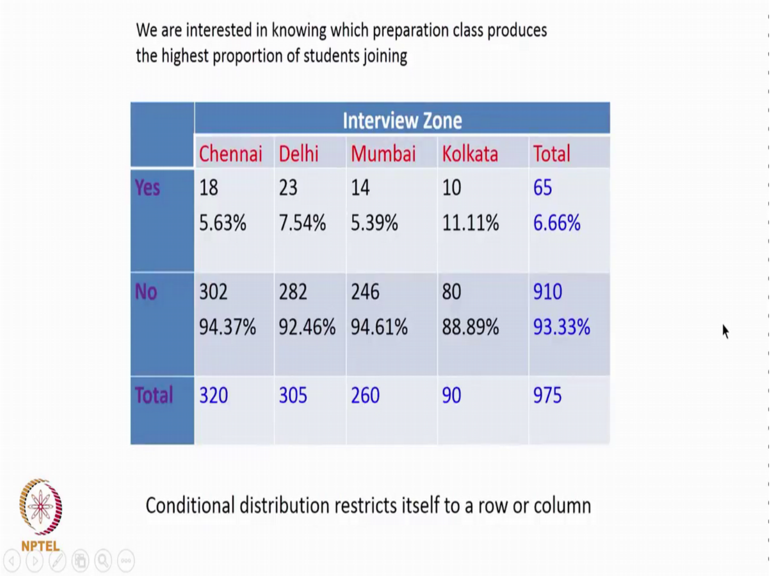
Conditional distribution restricts itself to a row or column

Now, we are in if we are interested in knowing is there an association between people joining and the class or we are interested in knowing which class produces the highest proportions of students joining. So, the overall proportion of students joining is 1.69 which is 65 divided by 3849. Now, this proportion is 37 out of 2166 joined which is 1.71, 18 out of 1047 joined which is 1.72 and 10 out of 636 joined which is 1.57. The average 65 out of 3849 joined which is 1.69 percent.

The distribution of a variable that is restricted to cases satisfying a condition is called a conditional distribution. So, in this case the condition is, yes, joining the program across this. So, the conditional distribution restricts itself to a row in this case the conditional variable is no and it again restricted itself to a row where 98.29 percent did not or could not joined out of those who went to class 1, 98.28 plus 2 and 98.43 to class 3. So, it restricts itself to a row.

Now, we can look at the other one out of class one those who went then they realize that there is a yes and there is a no. now, it restricts itself to a column where we say 1.71 percent could get in 98.29 percent did not. So, it restricts itself to a row or to a column.

(Refer Slide Time: 24:09)



We are interested in knowing which preparation class produces the highest proportion of students joining

	Interview Zone				
	Chennai	Delhi	Mumbai	Kolkata	Total
Yes	18 5.63%	23 7.54%	14 5.39%	10 11.11%	65 6.66%
No	302 94.37%	282 92.46%	246 94.61%	80 88.89%	910 93.33%
Total	320	305	260	90	975

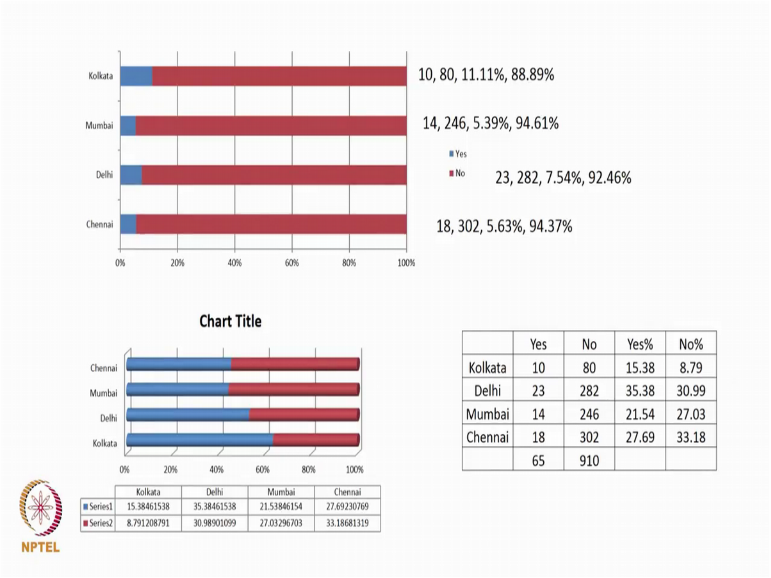
Conditional distribution restricts itself to a row or column

Now, we are interested again which one does the highest. Now, we change the contingency table to something else and we say that people who joined and people who did not join and then we also look at places where interviews happened.

So, now we look at the case where out of all those who were who had applied now certain number of people were called for interview and we say a total of 975 people were called for interview, out of which 65 people finally, joined and 910 did not or could not join the program. Now, we have data which is the 65. Now, the interview brings us another categorical variable and this categorical variable could be place of interview which could be Chennai, Delhi, Mumbai and Calcutta. So, we restrict ourselves to four places. So, one categorical variable is place of interview and the other categorical variable is yes or no.

So, now, we can find out an association and try to see whether is there an association where the city where the person was interviewed had a higher proportion or a more meaningful proportion or is there an association between the city and the selection. So, we can answer that question and we can do a similar analysis and these computations are shown. So, here the conditional distribution is the city with a yes or no it restricts itself to a column and yes or no, again with respect to the city with restricts itself to a row.

(Refer Slide Time: 25:45)



	Kolkata	Delhi	Mumbai	Chennai
🟦 Series1	15.38461538	35.38461538	21.53846154	27.69230769
🟥 Series2	8.791208791	30.98901099	27.03296703	33.18681319

	Yes	No	Yes%	No%
Kolkata	10	80	15.38	8.79
Delhi	23	282	35.38	30.99
Mumbai	14	246	21.54	27.03
Chennai	18	302	27.69	33.18
	65	910		

Now, we look at some pictorial representation of this data. So, the first picture this is a bar chart, but this bar chart also represents in the form of percentages and then it shows the four places where for example, the interviews were held and then it says that out of this 10 and 80, so, out of 90 people who let us say attended the interviews in Kolkata 10 people joined the program. So, 11.11 percent joined the program and 88.89 percent did not or could not join the program that is shown here the 11.11 percent is shown here in the blue color and the 88.89 percent of Kolkata is shown in the red color.

Similar charts are shown for Mumbai, Delhi and Chennai and assuming that these are the four cities where interviews were held. So, Mumbai the 5.39 percent who actually were called and attendant joined the program a 94.61 percent could not join or did not join or were not selected and so on. Similarly, we can see this graphs for Delhi and Chennai which are the four cities that we are looking.

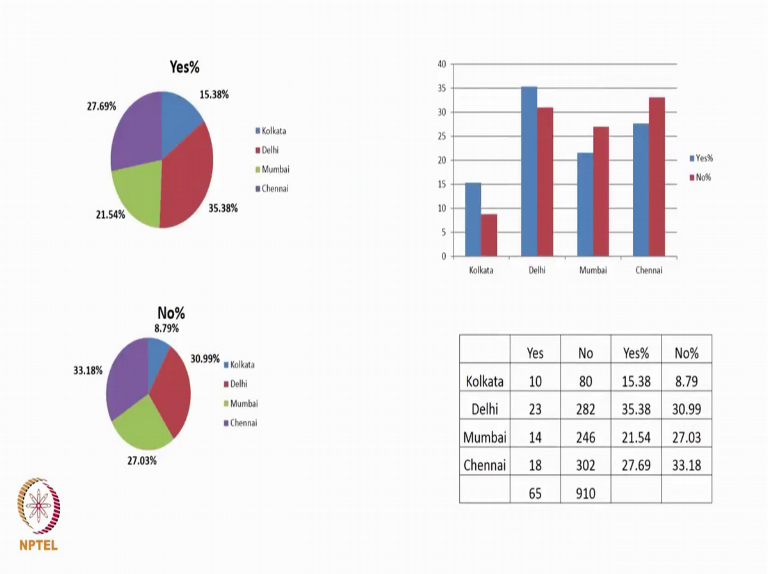
We also show another chart which you can generate using software and let me explain this chart. Now, let us look at Kolkata in this chart. Now, Kolkata 15 percent of the people now 65 people joined out of which 10 people were from Kolkata and 15 therefore, 15.38 percent of the people who joined are from Kolkata, 35.38 percent are from Delhi and so on. Now, among those who did not join or could not join or were not selected eighty are from Kolkata out of 910, which is 8.79 percent.

So, these percentages add up to 100, these also add up to 100, but now let us try to understand this picture. If you look at Kolkata 15 percent and 8 percent are here. So, this 15 percent is the blue which is here, the 8.79 is the red. Now, this length of this if assuming that the total length is 100 or 100 percent the length of this blue is actually 15.38 divided by 15.38 plus 8.79 and that comes to about 62 or 63 percent.

For Delhi it will the blue length of the blue color or blue part of the bar will be 35.38 divided by 35.38 plus 30.699 which would be just above 50 percent and you can see this here. For Mumbai you can see it is 21.54 divided by 21.54 plus 27.03 which will be less than 50 percent and you can see it.

So, this is another representation, but we have to understand what this bar chart is actually representing and this bar chart tries to tell that while 15.38 percent of the people who joined came from Kolkata and 8.79 percent of the people who did not or could not join came from Kolkata, the relative percentages of these are shown in the blue and red bars respectively.

(Refer Slide Time: 29:39)



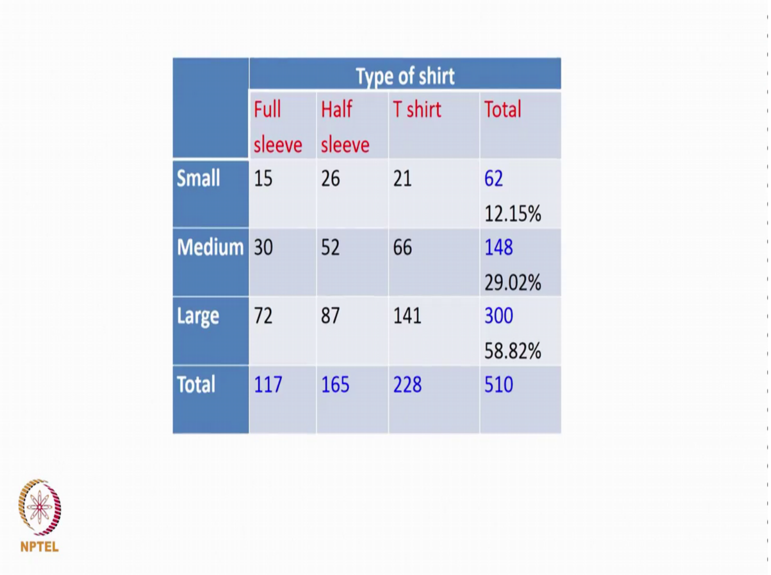
	Yes	No	Yes%	No%
Kolkata	10	80	15.38	8.79
Delhi	23	282	35.38	30.99
Mumbai	14	246	21.54	27.03
Chennai	18	302	27.69	33.18
	65	910		

We can show the same data in two different forms. Now, this is a bar chart representation straight away would say that if it is Kolkata then we are talking about 15.38 percent and 8.79 percent and you can see them in the blue bar as well as the red bar respectively. For Delhi it is 35.38 and 30.99 and so on.

Now, all the blues will add to 100 percent and all the reds will add to 100 percent. You can also think of another bar chart which is not shown here where the 100 percent of the blue is actually divided into four parts with 15.38 for Kolkata 35.38 for Delhi and so on. And, similarly the 100 percent for the red it also divided into four parts the Kolkata part which is 8.79, the Delhi part which is 30.99 and so on.

More simpler representation assuming that these are percentages and we want to generalize them and then say out of those who joined 15.38 came from Kolkata. So, the pie chart shows this representation and the pie chart also shows the percentages of people who did not or could not join the program from the four cities where they were interviewed.

(Refer Slide Time: 31:05)

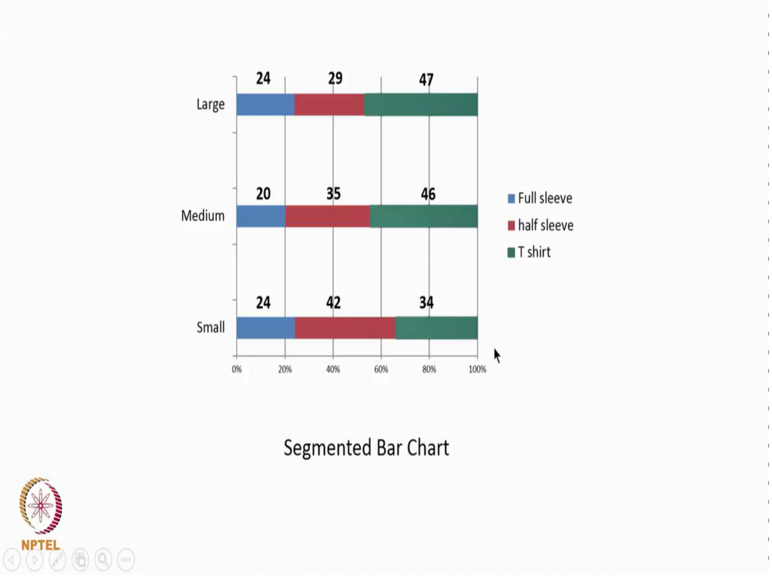


	Type of shirt			
	Full sleeve	Half sleeve	T shirt	Total
Small	15	26	21	62 12.15%
Medium	30	52	66	148 29.02%
Large	72	87	141	300 58.82%
Total	117	165	228	510

There can be another type of an example which could be the types of shirts. So another example of association between categorical variables so, we could have let us say about 510 T shirts were sold in small, medium and large three different sizes and in three different types which is a full sleeve, a half sleeve and a T shirt. So, there would be sleeve shirts full sleeve, half sleeve and T shirt with small, medium and large. One set of categorical variable is the size based which is small, medium and large, the other would be type of the shirt which is full sleeve, half sleeve and a T shirt.

In a similar manner using the data we can calculate these percentages and these are shown here.

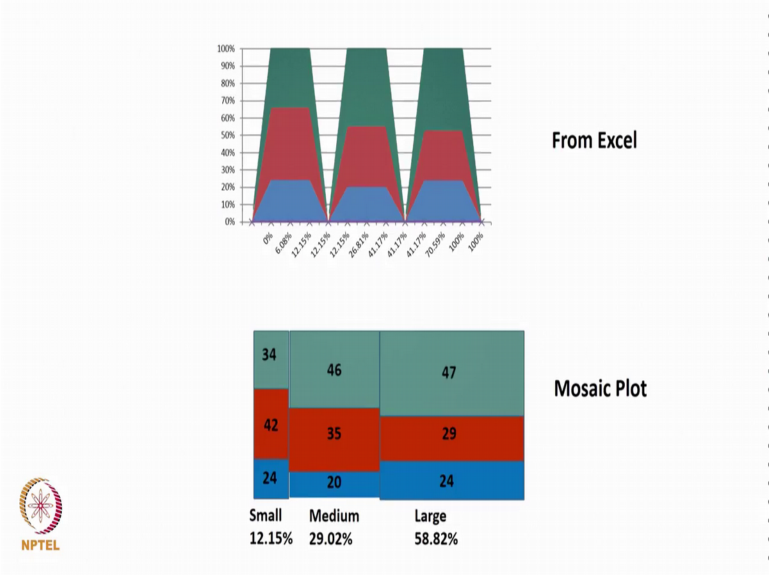
(Refer Slide Time: 31:55)



Segmented Bar Chart

We have also shown these in the form of a segmented bar chart with the numbers correspondingly for large medium and small for the three categories of shirts.

(Refer Slide Time: 32:04)

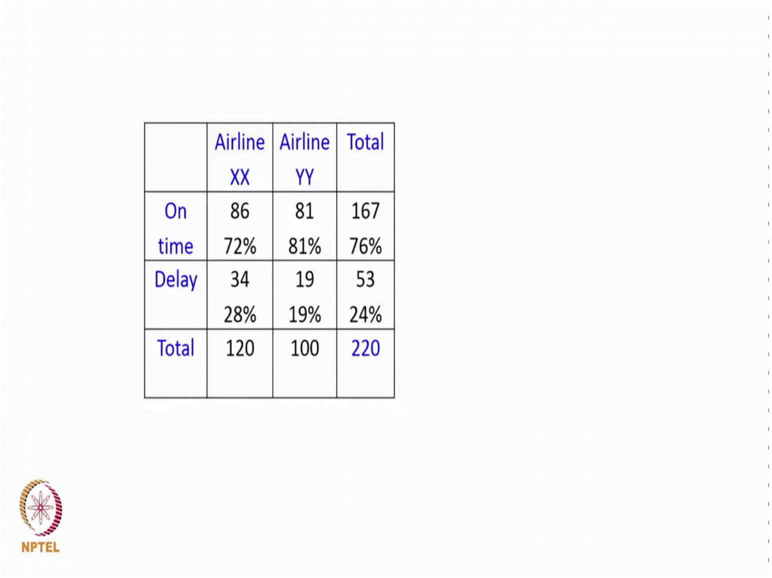


From Excel

Mosaic Plot

The same data can also be represented using mosaic plot which is shown here, these are all ways of representing this data which is earlier represented in the table.

(Refer Slide Time: 32:16)



	Airline XX	Airline YY	Total
On time	86 72%	81 81%	167 76%
Delay	34 28%	19 19%	53 24%
Total	120	100	220

We could look at another type of association; let us say we could think of two airlines which we call XX and YY and some data on on time arrival and a delay. So, there is one variable which is a time of arrival which is on time and delay and the other categorical variable would be airline XX and airline YY and we could think of 220 flights for which data has been taken and then we could get some numbers like 86, 81 and the corresponding percentages.

So, for this kind of a data we would be interested in finding out if there is an association between on time arrival and delay, but with different airlines that are under consideration. So, we look at data of about 220 flights and we can make a table like this. Now, how do we actually compute the measure? Is there a measure that we can use and come to a conclusion that there is association we see those things in the next lecture.