

# IIT Madras ONLINE DEGREE

# Statistics for Data Science - 1 Professor Usha Mohan Department of Management Studies Indian Institute of Technology, Madras Lecture 4.9

## Association between categorical and numerical variables

(Refer Slide Time:	0:17)	TE	
	Statistics for Data Science -1  L-Association between categorical and numerical variable		
	Introduction	Х, Ү	
INSZ	<ul> <li>Understand the association numerical variable.</li> <li>Assume the categorical var (dichotomous)</li> </ul>	between a categorical variable and iable has two categories  Gender Charle  Female  INTUMO HIGH  LOW	

So, in this portion we are going to understand how to capture the association between a numerical and a categorical variable. Here I am assuming my categorical variable has only two categories, in other words if I am looking at gender for example, it has two categories male and female. I could have another variable which is say income, I could just look at two categories which is just high category and low category.

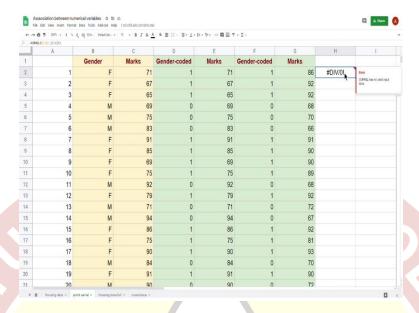
So, I am coming my categorical variable has two (variables) two categories and this is referred to as a dichotomous variable. So, now we are going to see how we summarize the association or understand the association between a categorical variable and a numerical variable. Let us look at the following example.

Statistics for Data Science -1 L—Association between categorical and numerical variable



Example 1: Gender versus marks





So, suppose a teacher was interested in knowing if female students perform better than the male students in her class. So, what does she collect? She collected data from 20 students. Now these 20 students and the marks obtained in the subject she is teaching on 100. A very small subset of the data set, so you can, you can see that there are 20 students. I will just show it, show the data to you.

So, I have the first student is a girl who has obtained 71 on 100, the second student is again a girl who has obtained 67 on 100, the sixth student is a male who has obtained 83 on 100, the twentieth student is a male who has obtained 90 on 100. So, these are the marks of 20 students and the data that is recorded is the gender. Now, gender is a categorical variable, now the levels of these categorical variables are two. The first level is a female, the second level is a male, it is nominal because there is no order in this categorical variable.

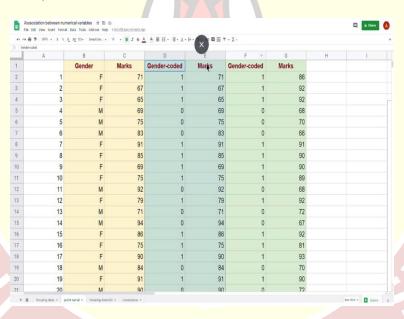
Whereas marks is a numerical variable, it is can take any value, here I have seen that 71 out of 100, 67 out of 100, this is a numerical variable. Now, we are interested in understanding weather females perform better than males, so we want to know what is the association between, first of all we are asking the question, is there an association between gender and the marks, or are they not associated with each other.

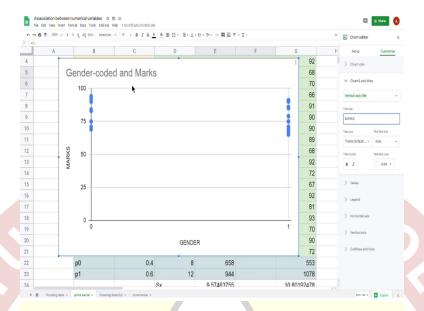
We introduced a concept which is referred to as a point bi-serial correlation measure. So, let us start by looking at a scatter of this data. To look at a scatter of this data what I first do is the

following, I code my categorical variable. What do I mean by coding my categorical variable? I have gender female and male here so I am just coding it, because now I am looking at whether I can quantify the correlation between these two variables. So if I just write for example, if I just try and see I will put a correlation measure and I, you can see that it returns an error because it says that correlation the (valid), it has no valid input data.

It has no valid input data because this is categorical and it is text. So, let me code this variable. So, this coding I am just arbitrarily choosing female as 1 and male as 0, so you can see that this data is coded in the following form. So, here I have two data sets, this could be the data obtained in one test and this could be the data or this could be the marks obtained in of the same 20 people in one test and this could be the marks obtained of the same 20 people in another test.

(Refer Slide Time: 4:52)

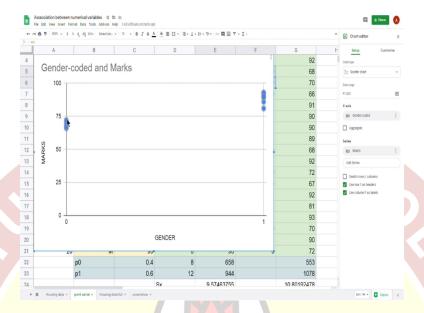




So, let us first start with a scatter plot of this data. So, again I go and I plot a scatter plot. I do not want a line chart, I want a scatter plot. And within the scatter plot I am using my column *D* as labels because I, column *D* is taking only two values, I am using column *D* as labels. And within this column *D* again what I do is, I go and I customize it further by looking at the major spacing time which is step and I am putting a step value of 1. So, you can see that my, I have constructed my scatter plot which, in which I have my *X*, I can go back to my titles, the title is gender coded versus marks.

My horizontal axis I have gender and on my vertical axis I have marks out of 100. And the way I can interpret this scatter plot is, you can see that this 0, 0 represents a male here and 1 represents a female here. So you can see that whether it is male or female all of them have obtained marks in the same range. So, the scatter plot tells us an important story here and the story which the scatter plot, this scatter plot tells us is irrespective of gender the distribution of marks seems to be the same, because both female and male seem to have performed equally well in a particular range. But now let us look at the second dataset.

(Refer Slide Time: 7:00)



So, if I go back and change this data range, let me look at the data range taking so, it is sorry F21 to F1 to G21 that is a second data range of my data. So, for this data range F1 to G21. Now, you see that there is some difference. What is difference you notice here? You can see that 0 or the males are clustered in this region and females are clustered in this region to indicate females have performed better, they have obtained higher marks than the males in general.

(Refer Slide Time: 7:47) 20:03, 22:11

Statistics for Data Science -1

L-Association between categorical and numerical variable

Example 1: Scatter plot

Gender-coded and Marks

100

130

130

140

Gender

Gender

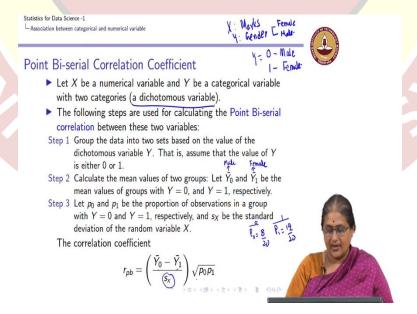
Gender

Example 1: Scatter plot



So, you see that in the earlier case, so you can see that in the earlier case that is in this case, the first case I see that 1 is again female. 0 is a male. And you can see that the marks obtained are the same. Whereas in the second case again I can see that female seem to have obtained higher marks than the males in general. This is a distribution of marks for males and the females. So, the question is, again here I do not have a line, I have just these points. Fitting a line for this kind of data is of no use to me, so how do I summarize the strength of this association.

(Refer Slide Time: 8:35)



## 4-0-Male 1- Female

### Point Bi-serial Correlation Coefficient

- ▶ Let *X* be a numerical variable and *Y* be a categorical variable with two categories (a dichotomous variable).
- ► The following steps are used for calculating the Point Bi-serial correlation between these two variables:
- Step 1 Group the data into two sets based on the value of the dichotomous variable Y. That is, assume that the value of Y is either 0 or 1.
- Step 3 Let  $p_0$  and  $p_1$  be the proportion of observations in a group with Y=0 and Y=1, respectively, and  $s_X$  be the standard deviation of the random variable X.

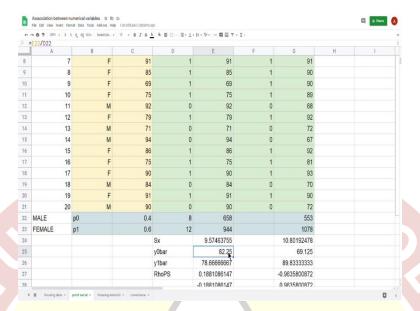
The correlation coefficient







				A ⊞ ⊞ - E - 1 - H	·· 》· □ □ □ 7	. • Σ ·			
=71	+67+65+91+85+6 A	9+75+79+86+75+9	00+91 C	D	e	r	G	Н	T.
1	Α	Gender	Marks	Gender-coded	Marks	Gender-coded	Marks	п	
2	-1	F	71	Gender-coded	71	Gender-Coded	Walks 86		
3	2	F	67		67			=71+67+65+91+8	54604754704964
-		-		1			92	5+90+91	31031731731801
4	3	F	65	1	65	1			
5	4	M	69	0	69	0	68		
6	5	M	75	0	75	0	70		
7	6	M	83	0	83	0	66		
8	7	F	91	1	91	1	91		
9	8	F	85	1	85	1	90		
10	9	F	69	1	69	1	90		
11	10	F	75	1	75	1	89		
12	11	M	92	0	92	0	68		
13	12	F	79	1	79	1	92		
14	13	M	71	0	71	0	72		
15	14	М	94	0	94	0	67		
16	15	F	86	1	86	1	92		
17	16	F	75	1	75	1	81		
18	17	F	90	1	90	1	93		
19	18	M	84	0	84	0	70		
20	19	F	91	4	91	1	90		
20	20	M	90	0	91	0	72		



For this we have a measure which we refer to as a point bi-serial correlation coefficient. So, let *X* be a numerical variable and *Y* be a categorical variable. In our example, *X* is my marks it is the numerical variable, *Y* is the gender which is the categorical variable. I am assuming a dichotomous variable, gender has two levels, again I have a level of female and male. How do I compute the point bi-serial correlation between these two variables? The first thing which we do is the following.

We go back, you group the data into two sets based on the value of the dichotomous variable *Y* that is I am saying *Y* takes a value 0 or 1. Here in my example I have assumed *Y* takes the value if the gender is male and *Y* takes the 1 if the gender is female. And that is how I have coded my data. So, you can go back and see that this is how we have coded the data here. So, if I have *Y*, it takes the value 1 if it is a female and 0 if it is a male. Group the data into two sets based on the value. So, this data again I have, so this data I have 1,71 this is the same data.

So, in this two sets I have two groups. What are the two groups? One is a female group and one is the male group. So, in the first step, in the second step you calculate the mean values of the two groups. How do I do this? One way to do this is, I can just find out what are the total number of females. So, if I count the total number of females it is 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 females. So out of 20, I have 12 females and 8 males. So, I have 8 males, so my 0, this is my male.

So, if I am summarizing it males, I have 8 males and I have 12 female students. So, the proportion is nothing but  $\frac{8}{20}$  which is going to be 0.4 and  $12 \mathbb{Z} 20 \mathbb{Z}$  which is 0.6. So, that is how and I am referring that to  $P_0$  and  $P_1$ . Now, what is the next step? The second step says, calculate the mean values of these two groups. So, how many females do I have? I have 12 females. What is the mean value?

It is going to be 71,71 + 67 + 65 +, so forth I can keep doing 91 + 85 + 69 + 75 + 79 + 86 + 75 + 90 + 91, which is 944. A simpler way to compute this is through using sum if function. How do you use a sum if function will be taught in the tutorial. It is a simpler (func) way that is if I have a female and just computing all those data points which correspond to the female, so I can use what is called the sum if function. So, this 944 represents the total marks obtained by female students.

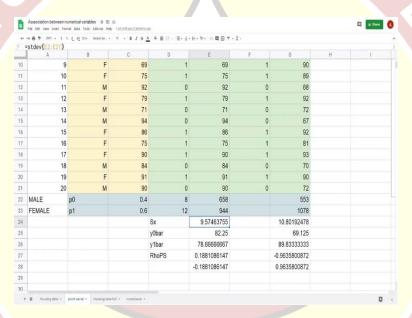
Similarly, 658 is the total marks obtained by all men students or male students put together. Now, what is the mean of this group? I have 12 students in this group, so the mean of this group which is going to be nothing but 944  $\boxed{2}12\boxed{2}$  which is equal to 78.66. So,  $\boxed{y_1}$  is giving me the mean of the female group. What is the average mark of the female students? Similarly the average mark of the male students is 82.25. So, if you go back to step two, it is saying that compute  $\boxed{y}\boxed{2}0\boxed{2}$  and  $\boxed{y_1}$  which are the mean values of the group.

So,  $y \ @ 0 \ @ 1$  here is the mean of the group of male students,  $\overline{y_1}$  is the mean of the group of female students. Then I can find out what is the proportion of observation. Again I know that  $P_0$  is going to be  $8 \ @ 20 \ @ 2, p_1 = \frac{12}{20}$ , this is the proportion of observations of the group which I am coding 0 and this is the group I am coding 1. Finally  $S_X$ ,  $S_X$ , X is my numerical, it is a single numerical variable which is the marks. Remember when I have numerical variable, I can define what I call a standard deviation measure which is telling me about the variability of that particular variable.

 observation and  $n_0$  is the number of observations in my 0th group, the group that is coded 0,  $n_1$  is a number of observations in the group that is coded 1.

Then n @ 0 @ 2 n @ = p @ 0 @,  $p_1$  is n @ 1 @ 2 n @. Some books or some authors instead of looking at this as  $p_0 \times p_1$  use a sample correction which is (the) only this term is replaced by  $n @ 0 @ 2 n - 1 @ \times n @ 1 @ 2 n @$ . So, you can see again this n-1 which keeps playing a role always when you talk about sample characteristics. So, this is how you compute a point bi-serial correlation coefficient of a data which has both numerical and categorical variable.

(Refer Slide Time: 17:19)



So, for this dataset again going back I have computed the point bi-serial coefficient which is . 188 and this for the next data set is -.96. Again is there a difference between . 188 and -.188. See, this 0 and 1 I could have change female to be 0 and male to be 1 then I would have got a different sign. That is what I have computed here, there is no, since this is, there is no order, there is no hard and fast rule that female should be 1 and male should be 0. I could have change the ordering.

(Refer Slide Time: 18:03)

1 -	67 20 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1										
Ī	if(B2="F",0,1) A	В	С	D		E		F	G	Н	1
1	0.0	Gender	Marks	ender-co	ded	Marks	G	ender-coded	Marks		
İ	1	F	71	= if(82="F	",0,1)		71	1	86		
	2	F	67	Df(logical_copress value_if_false)	ion, value_if_	trw. ^ X	67	1	92	944	
İ	3	F	65	EXAMPLE IF(A2 * "foo", "A2	to deat to	is not forth	65	1	92		
	4	М	69	ABOUT Setures one value if a lo it in FALSE.	olcai espression is:	TRIF and profes if	69	0	68		
	5	М	75	t is FALSE.	gue 109-1110111		75	0	70		
N.	6	М	83	logical_expression An expression or releve that represents some to	ceto a cell contain	nng ac-egresson	83	0	66		
	7	F	91	That represents some to vultar_if_true The value the function is	gcal value, i.e. 190 duma V inamal and	A OFFICE	91	1	91		
	8	F	85	value_if_false The value the function re	starre Finnical our	marrina ir EA/CE	85	1	90		
)	9	F		Learn more			69	1	90		
1	10	F	75		1		75	1	89		
2	11	М	92		0		92	0	68		
3	12	F	79		1		79	1	92		
1	13	М	71		0		71	0	72		
5	14	M	94		0		94	0	67		
5	15	F	86		1		86	1	92		
	16	F	75		1		75	1	81		
3	17	F	90		1		90	1	93		
)	18	М	84		0		84	0	70		
)	19	F	91		1		91	1	90		
٠											
+	Association between nur	M  oint serial - Housing date I  merical variables str (3) of the Data Tools Addions Help	2		n		90	n	72		Q a Share
+	Housing data - p  Association between nur ide Edit View Isset Forms  F 1 2005 - 1 5  ((E25-E26)/E24)	merical variables & D3 of Data Tools Address Help  o_ ao_ 133+  **sqrt((D22/20)*6	b last off war recording	n A   A   B   S -   S				Σ.			
+	Housing data - p  Association between num  ife Edit View Insert Forms  if 7 200% - d %  ((E25-E26)/E24)	merical variables of CB of the Data Tools Address Help of Sept. Se	D23/19))	n A   4. III II +	≅• 1• 1+•	- >- □ E	□ <b>∀</b> •	Σ.	G	Н	
+ 	Housing data - p  Association between nur ide Edit View Insert Forms  The 2005 - d to  ((E25-E26)/E24)  A  10	merical variables & D3 of the D4 of	b)	n A & B S - (	ē · ± · 1÷ ·		₩ ▼・	Σ · F 0	G 89		
+	Heusing data + p  Association between num  te Edit View Isset Forma  The P 2005 + 5 h  (E25-E26)/E24)  A  10  11	eint seral - Housing date I merical variables & ® ® ø it Data Tods Addees Help  9_ 82 83+ **sqrt((D22/28)*(  8 F M	bl - coverience -  1 interfit use unconticus 10 - B x & A  023/19))  C  75	n A & B S - (	■・±・トゥ・ 0 1		75 92	Σ. F 0	G 89 68		
+	Housing data - p  Association between num  ie Edi View Issert Forms  A  10  11  12	merical variables of DI of DI of Tools Address Method Tools Address Method Tools Address Method Tools Address Method Tools Address Method Tools Address Method Tools Address Method Tools Tools Address Method Tools Too	1 octoff assumenting C	n A & B S - (	■・ 土・ I÷・ O 1 0		75 92 79	Σ - F 0 1 0 0	6 89 68 92		
+	Housing data - P  Association between num  te Edit View Issert Forma  (E25-E26)/E24)  A  10  11  12  13	orit serial   Hoosing data   Hoosing data   to the transmission of the transmission of the   Barrier   Barrier   M  F  M	10 - Covertence - 1 Not reft use use control to 10 - 8 r 4 / (D23/19)) C 75 92 79 71	n A & B S - (	回・土・ト;・ 0 1 0 1		75 92 79 71	Σ. F 0 1 0	6 89 68 92 72		
+	Nassociation between nurrele Edit Vere Heart Forms  (E 256 Vere Heart Forms  (E 25 - E26)/E24)  A  10  11  12  13	merical variables or to a tour tour tour tour tour tour tour tour	D23/19)) C 75 92 79 94	n A & B S - (	0 1 0 1		75 92 79 71 94	Σ.  F 0 1 0 1 1 1	89 68 92 72 67		
+ = (	Nassociation between nurrelection to the second sec	merical variables of 00 of the Trans Address May 121 121 121 121 121 121 121 121 121 12	all - coverience  all antiditives overcommunity of the series of the	n A   4- B E +   5	0 1 0 1 1 0		75 92 79 71 94 86	Σ. F 0 1 0 1 1 1	89 68 92 72 67 92		
+	Heaving data = p   Association between num   te Edit View Isset Feeta   The P   20% - 5   5     The P   20% - 1   10     The P   20% - 1     The	merical variables of 00 of but Trook Address Mely 2, 20 101+  **sqrt((022/20)**/  **B  F  M  M  M  F  F  F	Lat - covertience - Lat - covertience - Lat - covertience - Lat - covertience - Lat - covertience - Covertience - Lat - covertience - Covertie	n A   4- B E +   5	0 1 0 1 1 0 0		75 92 79 71 94 86 75	F 0 1 1 0 1 1 1 0 0 0 0	89 68 92 72 67 92 81		
+ =(1 11 22 33 44 55 77 8		merical variables of 00 or total total address with the property of the total address with the property of the total address with the property of the property	Lat - covertience - Lat - covertience - Lat - covertience - Lat -	n A   4- B E +   5	0 1 0 1 1 0 0 0		75 92 79 71 94 86 75 90	F 0 1 1 0 1 1 1 0 0 0 0 0 0	89 68 92 72 67 92 81		
+ = (	Makesing data -   p	merical veriables or 00 or bos Total Address Melpolitics or 00 or bos Total Address Melpolitics or 00 or bos Total Address Melpolitics Mel	Constitutes - Constitutes	n A   4- B E +   5	0 1 0 1 1 0 0 0 1		75 92 79 71 94 86 75 90 84	F 0 1 1 0 0 1 1 1 0 0 0 0 1 1 1	89 68 92 72 67 92 81 93 70		
+ =(11 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8 8 9 9 0 0	Makeing data -   P   Makeing	merical veriables of the control of	11 10 CONSTITUTES - 1 10 11 11 11 11 11 11 11 11 11 11 11 1		0 1 0 1 1 0 0 0 1 0 0		75 92 79 71 94 86 75 90 84 91	F 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0	89 68 92 72 67 92 81 93 70		
+ = ( 1 = ( 1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8 9 9 9 1 1 1	Makesing data -   P   Makesing data -   P	merical veriables of the control veriables of	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	M (本) 田 昭 +   1	0 1 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 0	Ε	75 92 79 71 94 86 75 90 84 91	F 0 1 1 0 0 1 1 1 0 0 0 0 1 1 1	89 68 92 72 67 92 81 93 70 90		
+	Resolution   Page   Resolution   Resolutio	merical variables of 00 of the 10 of	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	n A A B E - D	0 1 0 0 1 1 0 0 0 1 1 0 0 0 1 1 1 1 1 2 1 2	E	75 92 79 71 94 86 75 90 84 91 90	F 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0	89 68 92 72 67 92 81 93 70 90 72		
+ 1 2 1 1 1 2 2 3 3 4 4 5 5 6 6 6 7 7 8 8 8 9 9 0 0 1 1 2 2 1 3 3 1 1	Resolution   Page   Resolution   Resolutio	merical veriables of the control veriables of	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	n A A B E -	0 1 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 0	E	75 92 79 71 94 86 75 90 84 91 90 90 844 86558	F 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0	89 68 92 72 67 92 81 93 70 90 72 1078		
+     = (1   1   2   2   3   4   4   5   5   6   6   7   7   8   9   9   1   1   2   1   3   3   4   4   4   6   6   6   6   7   7   8   9   9   9   9   9   9   9   9   9	Resolution   Page   Resolution   Resolutio	merical variables of 00 of the 10 of	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	n A   4. III 52 +   5	0 1 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1	9.57463	75 92 79 71 94 86 75 90 84 91 90 90 94	F 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0	89 68 92 72 67 92 81 93 70 90 72 1078 553		
=( 11 12 22 33 44 45 55 66 67 77 88 89 99 00 11 12 22 13 14 44 44 44 45 45 46 46 46 46 46 46 46 46 46 46 46 46 46	Resolution   Page   Resolution   Resolutio	merical variables of 00 of the 10 of	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	n A   4. ⊞ 53 +   5 D	0 1 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1	9.57463 78.66666	75 92 79 71 94 86 75 90 84 91 90 90 844 91 95 96 97 95 96 96 97 96 96 96 97 96 96 96 96 96 96 96 96 96 96 96 96 96	F 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0	89 68 92 72 67 92 81 93 70 90 72 1078 553 10.80192478 89.83333333		
+	Resolution   Page   Resolution   Resolutio	merical variables of 00 of the 10 of	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Sx y0bar y1bar	0 1 0 1 1 0 0 0 1 1 0 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 0 1 1 0 1 1 1 1 1 0 1	9.57463 78.66666 83	75 92 79 71 94 86 75 90 84 91 90 90 84 91 90 25 667 755 667	F 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0	89 68 92 72 67 92 81 93 70 90 72 1078 553 10.80192478 89.83333333 69.125		
+ + + + + + + + + + + + + + + + + + +	Resolution   Page   Resolution   Resolutio	merical variables of 00 of the 10 of	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	n A   4. ⊞ 53 +   5 D	0 1 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 1 1 1 1 1 1 2 1 1 1 1	9.57463 78.66666 8;	75 92 79 71 94 86 75 90 84 91 90 94 858 8755 667 2.25	F 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0	89 68 92 72 67 92 81 93 70 90 72 1078 553 10.80192478 89.8333333 69.125 0.9635800872		
+     = (	Resolution   Page   Resolution   Resolutio	merical variables of 00 of the 10 of	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Sx y0bar y1bar	0 1 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 1 1 1 1 1 1 2 1 1 1 1	9.57463 78.66666 83	75 92 79 71 94 86 75 90 84 91 90 94 858 8755 667 2.25	F 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0	89 68 92 72 67 92 81 93 70 90 72 1078 553 10.80192478 89.83333333 69.125		
=======================================	Resolution   Page   Resolution   Resolutio	merical variables of 00 of the 10 of	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Sx y0bar y1bar	0 1 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 1 1 1 1 1 1 2 1 1 1 1	9.57463 78.66666 8;	75 92 79 71 94 86 75 90 84 91 90 94 858 8755 667 2.25	F 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0	89 68 92 72 67 92 81 93 70 90 72 1078 553 10.80192478 89.8333333 69.125 0.9635800872		

So, I could have changed the ordering to be if it is female it is 0 and it is 1 here, otherwise. So, you can see that there is a flip in my this one, I have 12 males. So now this would become a female, because I had 12 females, sorry this would become a female now. This would be a male. 8 out of 12, this has, the standard deviation remains the same because I have not changed anything. My  $\overline{y_0}$  now is going to be your E22 by D22, y 2122 is going to be  $\frac{658}{8}$ .

Now, you can see that the point bi-serial which was earlier . 188 is now -.188. So depending on what is your coding, you are going to get this point bi-serial correlation coefficient that is the key difference between what I wanted to, what I wanted you to observe between numerical data, I have

a strong positive, or a strong negative, here it depends on how you are looking at the coding and that decides on what is your point bi-serial correlation coefficient.

So, whether it is a very, so here I can, so basically in absolute terms this would lie between 0 and 1 and it is closer to 0, it says that in this dataset my gender really has no association with the marks. And that is sort of validated by my scatter plot also for my first example which said that irrespective of which gender you are from, your marks seem to be distributed in the same way.

(Refer Slide Time: 20:15)

	File Edit View Insert For									
	ペラマ 2005 · 1 1 010	0 00 123+ Delauf (An. +	10 - B I & A	<u> </u>	H - 7 - ∞ ■ ■ T	* Σ -				
	A	В	C	D	E	F	G	Н	1	
		Gender	Marks	Gender-coded	Marks	Gender-coded	Marks			
	1	F	71	0	71	0	86			
	2	F	67	0	67	0	7 92	944		
	3	F	65	0	65	0	92			
	4	М	69	1	69	1	68			
	5	М	75	1	75	1	70			
	6	М	83	1	83	1	66			
	7	F	91	0	91	0	91			
	8	F	85	0	85	0	90			
	9	F	69	0	69	0	90			
	10	F	75	0	75	0	89			
	11	М	92	1	92	1	68			
	12	F	79	0	79	0	92			
	13		71	1	71	1	72			
	14		94	1	94	1	67			
	15		86	0	86	0	92			
	16		75	0	75	0	81			
	17		90	0	90	0	93			
	18		84	1	84	1	70			
	19		91	0	91	0	90			
+	Association between n	point serial - Mousing data  umerical variables \$2 G3  nat Data Tools Address Help	۵	1	90	1	72		Q a Share	
	Heusing data -  Association between n  File Edit View Issert For	point serial + Housing data	full = coverience =  O Just will war verseds up				72			
	Association between m  File Edit View Insert For  File 2005 + 3 11  Correct(12-12), (2-22)	point serial = Housing date  umerical variables dr 03  set Data Toda Addots Help  b 0 80 103+ Detait An. =	full - covertence -  Duct offt was percenture  10 - B I & I	D D	⊬- <b>ን</b> - ∞ <b>ቯ⊡</b> ₹	· · ː ·	G	Н		
	Association between nr Fife Edit View Issert For Pr 2005 - 8 1 correct(12:103, (2:03))	point sarial - Housing data umerical variables dr GI mat Data Toda Addons Help b. Q. 82 123 - Detait (Ac	full - coverlence -	D 1	H-> ∞ ■ ⊡ ▼	'+Σ. F	G 90			
	Hausing data →  Association between m     File Edit View Issert For	point sarial + Housing data  umerical variables st 03  mat Data Tools Addons Helpi  B  F	hall = coverlance =	D D	H->- □ ■ □ ▼ E 85	F 1 1	90 90			
+	Association between nr Fife Edit View Issert For Pr 2005 - 8 1 correct(12:103, (2:03))	point sarial + Housing data  ummerical variables st 03  mat Data Tools Addons Help  B  F	full - coverlence -	D 1 1 1	H-> ∞ ■ ⊡ ▼	'+Σ. F	G 90			
	Hausing data →  Association between m     File Edit View Issert For	point serial = Housing data umerical variables 2 to  and that Stock Address Help b, 0, 00 to Helself An  B F F M	hall = coverlance =	D 1 1	H->- □ ■ □ ▼ E 85	F 1 1	90 90			
	Association between mr fee toti Veer isset for the fee toti Veer isset for A P 200 × 8 1 Correlor A A 8 9 100 111 111 112	point serial = Housing data umerical variables : 2: 01 ent that 1 this Address Help h. 4, 49, 101 = Helet Mrs. +  F  F  M  F	hull - covariance +	D 1 1 1 0 0 1 1	H··≫· □■□▼ E  85 69 75 92	F 1 1 1 1 0 0 1 1	90 90 90 89 68 92			
	Association between me fee toti View isset for the fee fee fee fee fee fee fee fee fee f	point serial - Wousing data umerical variables or 10 nst Data Toda Addoors Help B F F M M F	©   1 statistical variations   1 statistical variations   1 statistical variations   10   8 x ← 4   C   S5   69   75   92   79   71	D 1 1 1 0 0 1 0 0	E 85 69 75 92 79 71	F 1 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0	90 90 99 68 92 72			
	B Housing data - Association between m File Edit View Issert Fer Committee Control (1975) 2019 - 8 8 9 100 111 11 12 13 13 14	point serial + Rossing data umanical variables & D umanical variable	©   101.01 and or reference   1   10   1   8   7   6   6   9   75   92   79   71   94	D 1 1 1 0 0 1 0 0 0 0	#+ * - = = = = = = = = = = = = = = = = = =	F 1 1 1 1 0 0 1 1 0 0 0	90 90 99 68 92 72			
	Association between fire 60 vew least for 60 vew least fo	point serial + Receiving data umerical variables or (b) at Data 1 bits 3 bits Address into part B  F  F  M  F  M  M  F	© 1 1 of all an according to 10 10 10 10 10 10 10 10 10 10 10 10 10	D 1 1 1 0 0 0 0 0 1 1	E 85 69 75 92 79 71 94 86	F 1 1 1 0 0 1 1 0 0 0 1 1	90 90 89 68 92 72 67			
1	## Shouling data - Association between fire fact view last fire fact view last fire fact view last fire fact view last fire fact view last fire fact view last fire fact view last fire fact view last fire fact fact fact fact fact fact fact fact	point serial - Woulding data umerical variables do 10 ns Des Total Address Indjud B F F M M F M M F F F	© 1 1 of all assurements as 1 1 of all assurements as 1 1 of all assurements as 1 1 of all assurements as 1 1 of all assurements as 1 1 of all assurements as 1 1 of all assurements as 1 of all assur	D 1 1 1 0 0 1 0 0 0 0	E 85 69 75 92 79 71 944 866 75	F 1 1 1 1 0 0 1 1 0 0 0 1 1 1 1 1	90 90 89 68 92 72 67 92 81			
		point serial - Wouling data umerical variables de 13 ns Des Toda Addoor Indjud  B  F  F  M  M  F  M  F  F  F  F  F	□ I and Left an	D	E 85 69 75 92 79 71 94 866 75 90	F 1 1 1 0 0 1 1 0 0 1 1 1 1 1 1 1 1 1 1	90 90 89 68 92 72 67 92 81			
!!!!!!!		point serial - Woosing data umerical variables dr 13 mr Data Toda Addoors Indyn B F F M M F M M F M M M M F M M M M M M		D 1 1 1 0 0 1 1 1 1 1 0 0 0 1 1 1 1 1 0 0 0 0 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 0 0 0 0 1 1 1 1 1 1 0	E 855 699 75 92 79 71 94 86 755 90 884	F 1 1 1 0 0 0 1 1 1 1 1 0 0 0 0 1 1 1 1	90 90 89 68 92 72 67 92 81 93			
		point serial + Rossing data umanical variables + 0: 0: en to so solos Address Help A	Companies   Comp	D 1 1 1 0 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 1 0 0 1	E 85 69 75 92 79 94 86 75 90 84 91	F 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 1	90 90 89 68 92 72 67 92 81 93 70			
		point serial - Resulting data umerical variables of 10 are the situation and serial se	** Constitute an American **  ** Institute an American **  ** B 2 * 6 4  ** B 2 * 6  ** B 2	D 1 1 1 0 0 0 1 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 1	E 85 69 75 92 79 711 94 86 75 90 84 91 90 90	F 1 1 1 0 0 0 1 1 1 1 1 0 0 0 0 1 1 1 1	90 90 89 68 92 72 67 92 81 93 70 90			
+		point serial + Receiving data umerical variables or (b) and bias 1 bias 1 bias 4 direct individuals of (b) 4 direc	Test   Constitutes   Constit	D	E 85 69 75 92 79 71 94 86 75 90 84 91 90 6658	F 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 1	90 90 99 68 92 72 67 92 81 93 70 90 72			
•		point serial - Resulting data umerical variables of 10 are the situation and serial se	** Constitute an American **  ** Institute an American **  ** B 2 * 6 4  ** B 2 * 6  ** B 2	D 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 0 1 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 0 1	E 85 69 75 92 79 71 94 86 75 90 84 91 90 6658 944	F 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 1	90 90 89 68 92 72 67 92 81 93 70 90 72 553			
•		point serial + Receiving data umerical variables or (b) and bias 1 bias 1 bias 4 direct individuals of (b) 4 direc	Test   Constitutes   Constit	D 1 1 1 0 0 1 1 1 0 0 1 1 0 0 8 8 12 Sx	E 85 69 75 92 79 91 94 866 75 90 84 91 90 6688 944 9.57463755	F 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 1	90 90 89 68 92 72 67 92 81 93 70 90 72 553 1078			
•		point serial + Receiving data umerical variables or (b) and bias 1 bias 1 bias 4 direct individuals of (b) 4 direc	Test   Constitutes   Constit	D 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 0 1 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 0 1	E 85 69 75 92 79 90 84 91 90 6688 9444 9.57463755 82.25	F 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 1	90 90 89 68 92 72 67 92 81 93 70 90 72 553 1078 10.80192478 69.125			
		point serial + Receiving data umerical variables or (b) and bias 1 bias 1 bias 4 direct individuals of (b) 4 direc	Tell   Constitutes   Constit	D 1 1 1 0 0 1 1 1 0 0 1 1 0 0 8 8 12 Sx	E 85 69 75 92 79 91 94 866 75 90 84 91 90 6688 944 9.57463755	F 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 1	90 90 89 68 92 72 67 92 81 93 70 90 72 553 1078 10.80192478 69 125 89.83333333			
		point serial + Receiving data umerical variables or (b) and bias 1 bias 1 bias 4 direct individuals of (b) 4 direc	Tell   Constitutes   Constit	D 1 1 1 0 0 1 1 1 1 0 0 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 1 0 0 1	E 85 69 75 92 79 711 94 86 75 90 84 91 90 658 944 9.57463755 82.25 78.66666667 0.1881086147	F 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 1	90 90 90 89 68 92 72 67 92 81 93 70 90 72 553 1078 10.80192478 69.125 89.833333333			
+		point serial + Receiving data umerical variables or (b) and bias 1 bias 1 bias 4 direct individuals of (b) 4 direc	Tell   Constitutes   Constit	D 1 1 1 0 0 0 1 1 1 0 0 0 8 8 12 Sx y0bar y1bar	E 85 69 75 92 79 71 94 86 75 90 84 91 90 658 944 9.5746375 82 25 78.66666667	F 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 1	90 90 89 68 92 72 67 92 81 93 70 90 72 553 1078 10.80192478 69 125 89.83333333			

Now, if we go to the second dataset, which is this F and G columns and do the same exercise for this dataset. Let me revert to my earlier coding of female is, does not make a difference but I just wanted to revert to my earlier coding and compute the point bi-serial. So, now again there are couple of things which I want you to see here. You can see that the total marks obtained by females here is way high than the total marks obtained by males, whereas here it was 944,658, here it is much higher.

If you look at the group means, the means in the first example were close to each other, the total mean of the male group was about 82 and the female was about 78, in fact males had a higher mean that the females here. Whereas here you can see that the males are, the female mean is about close to 90 whereas the male mean is close to 70, there is a huge difference, here there is not too much of a difference.

The standard deviation is again not very different this is 9.5 and 10.8 they are not very different from each other, but when you look at the point bi-serial correlation measure, this is very much close to one, in fact it is .963. Again that is something which is validated here by saying that yes the gender seems to play a role when I am seeking the association between gender and marks in this example.

So, we stop here. What we have seen in this week is understand about association between variables. We looked at association between categorical variables, we looked at association between numerical variables and we looked at association between a numerical and a categorical variable. Here we just restricted our attention to a categorical variable which has only two categories and we introduced what is called the point bi-serial correlation coefficient.

स्पिद्धिर्भवति कर्मजा