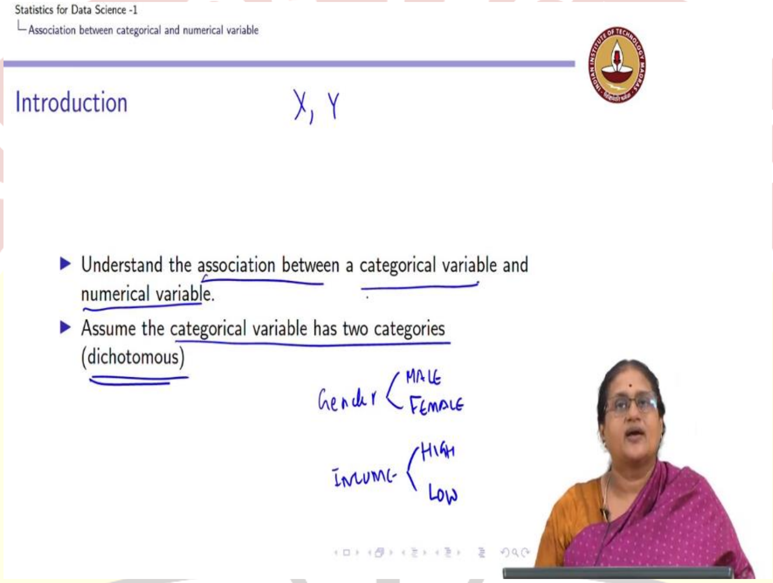
INDIAN INSTITUTE OF TECHNOLOGY MADRAS • सिद्धिर्भवति कर्मजा •

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)



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

Introduction

X, Y

- Understand the association between a categorical variable and numerical variable.
- Assume the categorical variable has two categories (dichotomous)

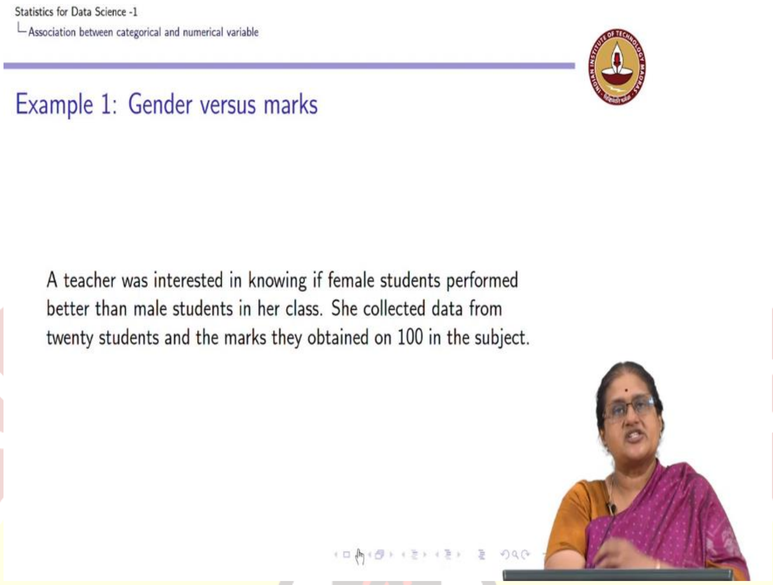
Gender { MALE, FEMALE

INCOME { 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.

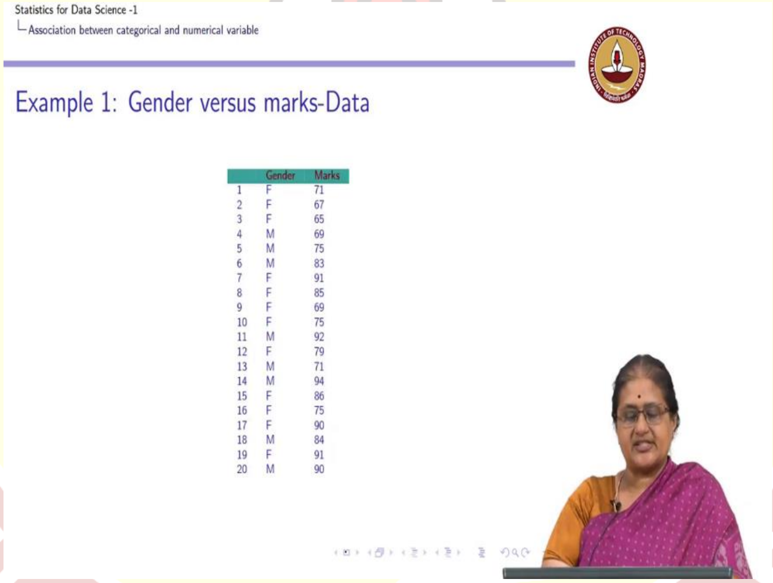
(Refer Slide Time: 1:17)



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

Example 1: Gender versus marks

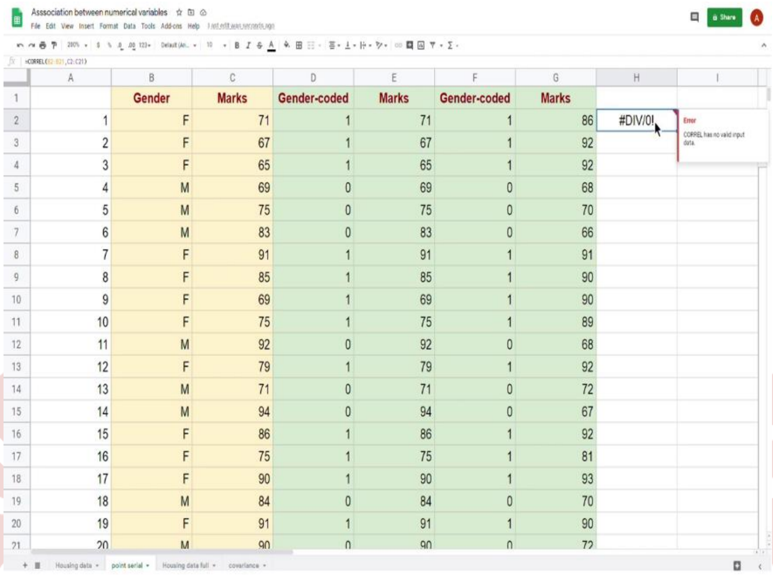
A teacher was interested in knowing if female students performed better than male students in her class. She collected data from twenty students and the marks they obtained on 100 in the subject.



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

Example 1: Gender versus marks-Data

	Gender	Marks
1	F	71
2	F	67
3	F	65
4	M	69
5	M	75
6	M	83
7	F	91
8	F	85
9	F	69
10	F	75
11	M	92
12	F	79
13	M	71
14	M	94
15	F	86
16	F	75
17	F	90
18	M	84
19	F	91
20	M	90


Asssociation between numerical variables
File Edit View Insert Format Data Tools Add-ons Help
Share
Housing data
point serial
Housing data full
covariance

=CORREL(B2:B21,C2:C21)

	A	B	C	D	E	F	G	H
1		Gender	Marks	Gender-coded	Marks	Gender-coded	Marks	
2	1	F	71	1	71	1	86	#DIV/0!
3	2	F	67	1	67	1	92	
4	3	F	65	1	65	1	92	
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	M	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	1	91	1	90	
21	20	M	90	0	90	0	72	

Error
CORREL has no valid input data.

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.

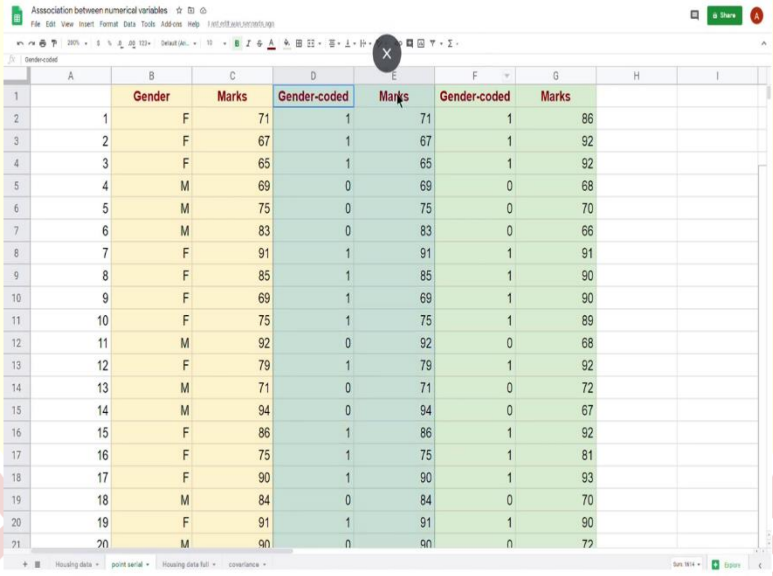
So, I have the first student is a girl who has obtained 71 on 100, the second student is again a girl who has obtained 67 on 100, the sixth student is a male who has obtained 83 on 100, the twentieth student is a male who has obtained 90 on 100. So, these are the marks of 20 students and the data that 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.

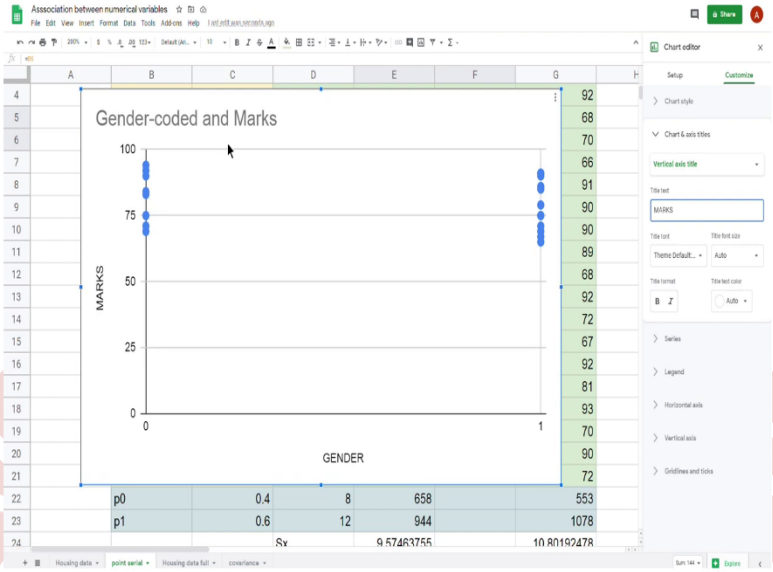
We introduced a concept which is referred to as a point bi-serial correlation measure. So, let us start by looking at a scatter of this data. To look at a scatter of this data what I first do is the following, I code my categorical variable. What do I mean by coding my categorical variable? I have gender 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)


Asssociation between numerical variables
File Edit View Insert Format Data Tools Add-ons Help
Share
Gender-coded
Housing data
point serial
Housing data full
covariance
Explore

	A	B	C	D	E	F	G
1		Gender	Marks	Gender-coded	Marks	Gender-coded	Marks
2	1	F	71	1	71	1	86
3	2	F	67	1	67	1	92
4	3	F	65	1	65	1	92
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	M	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	1	91	1	90
21	20	M	90	0	90	0	72



Vertical axis title

Title text: MARKS

Title font: Theme Default

Title font size: Auto

Title format: B I

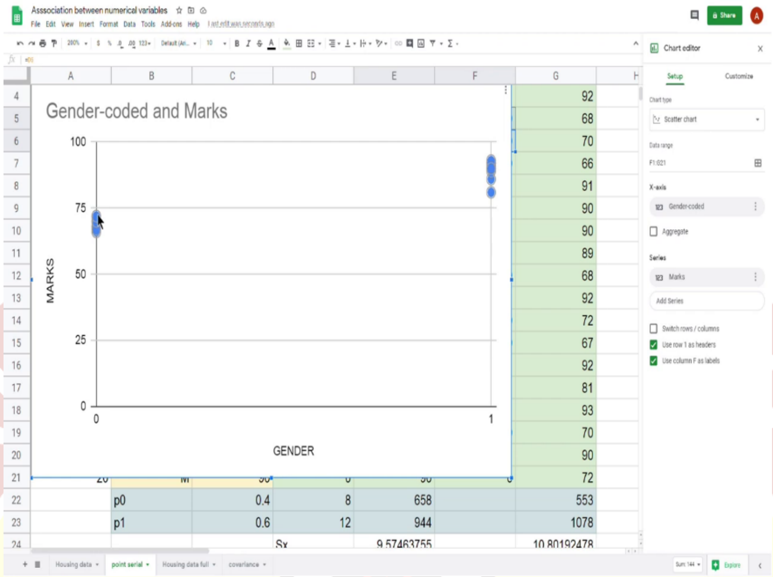
Title text color: Auto

	A	B	C	D	E	F	G
4							92
5							68
6							70
7							66
8							91
9							90
10							90
11							89
12							68
13							92
14							72
15							67
16							92
17							81
18							93
19							70
20							90
21							72
22		p0	0.4	8	658		553
23		p1	0.6	12	944		1078
24				Sx	9.57463755		10.80192478

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)



Asssociation between numerical variables

File Edit View Insert Format Data Tools Add-ons Help Last edit was seconds ago

Share

Chart editor

Setup Customize

Chart type

Scatter chart

Data range

F1:G21

X-axis

Gender-coded

□ Aggregate

Series

Marks

Add Series

□ Switch rows / columns

☑ Use row 1 as headers

☑ Use column F as labels

	A	B	C	D	E	F	G
4							92
5							68
6							70
7							66
8							91
9							90
10							90
11							89
12							68
13							92
14							72
15							67
16							92
17							81
18							93
19							70
20							90
21	20	M	90	0	90	0	72
22		p0	0.4	8	658		553
23		p1	0.6	12	944		1078
24			Sx		9.57463755		10.80192478

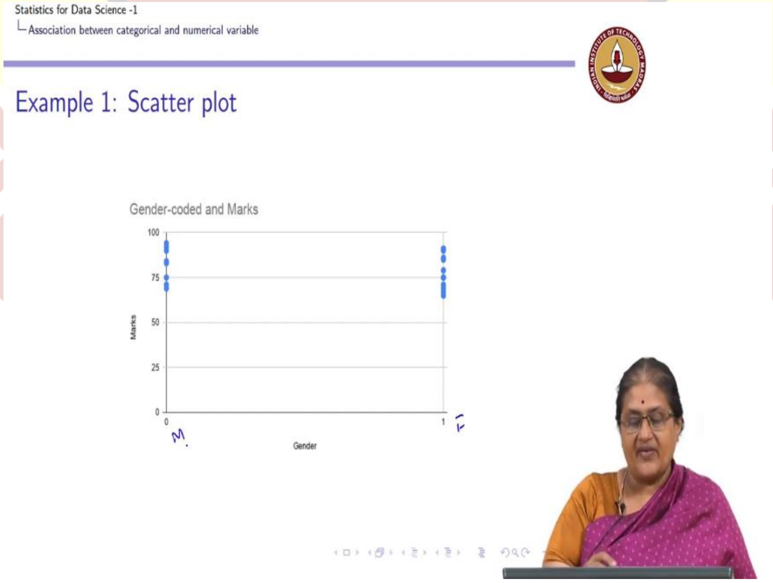
Housing data · point serial · Housing data full · covariance

Sum: 144

Explore

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.

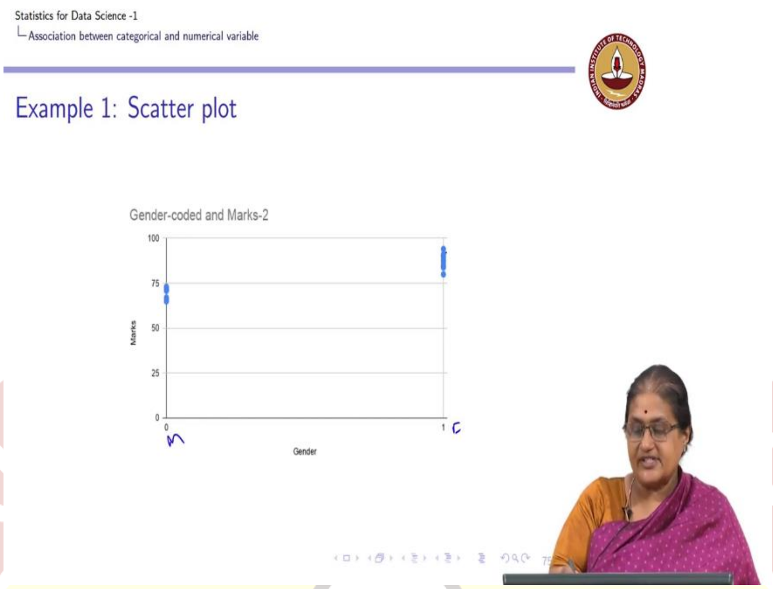
(Refer Slide Time: 7:47) 20:03, 22:11



Statistics for Data Science -1

└Association between categorical and numerical variable

Example 1: Scatter plot

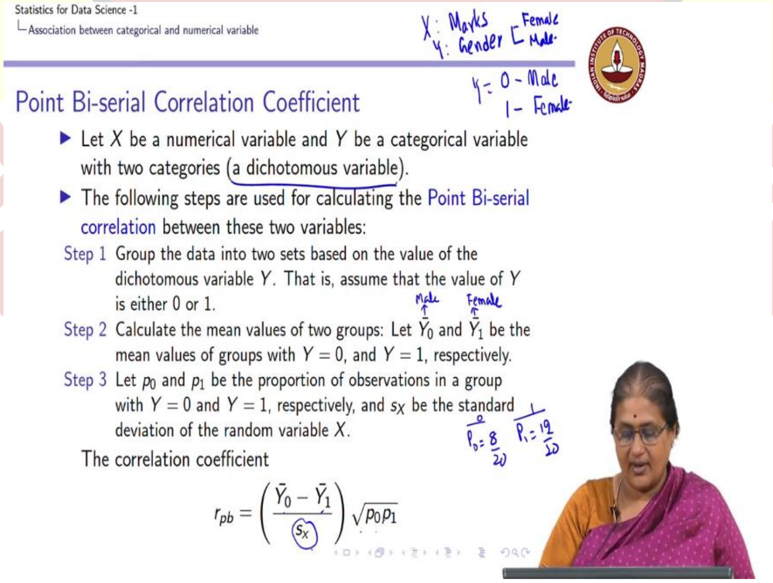


Statistics for Data Science -1
└ Association between categorical and numerical variable

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)



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

X : Marks
Y : Gender [Female, Male

Y = 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.

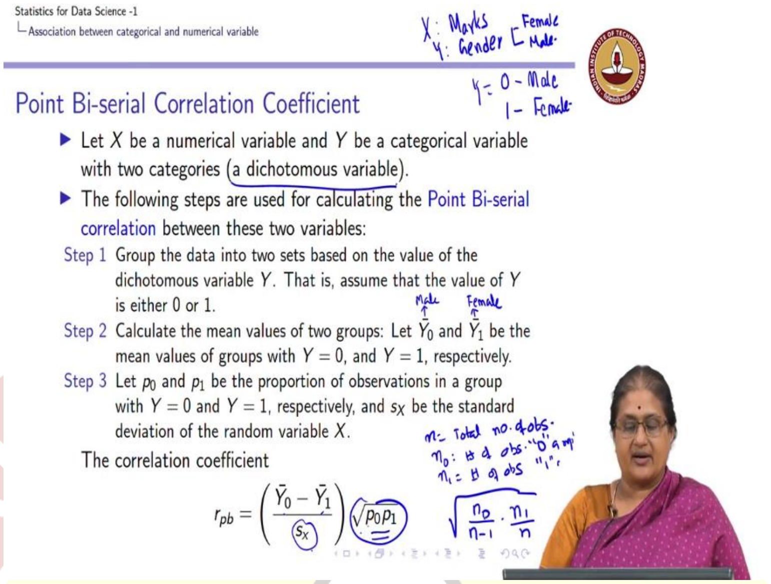
Step 2 Calculate the mean values of two groups: Let $\overset{\overset{\text{Male}}{\uparrow}}{\bar{Y}_0}$ and $\overset{\overset{\text{Female}}{\uparrow}}{\bar{Y}_1}$ be the mean values of groups with $Y = 0$, and $Y = 1$, respectively.

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.

$p_0 = \frac{8}{20}$ $p_1 = \frac{12}{20}$

The correlation coefficient

$$r_{pb} = \left(\frac{\bar{Y}_0 - \bar{Y}_1}{s_X}\right)\sqrt{p_0 p_1}$$



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

X: Marks
Y: Gender [Female
Male

Y = 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.

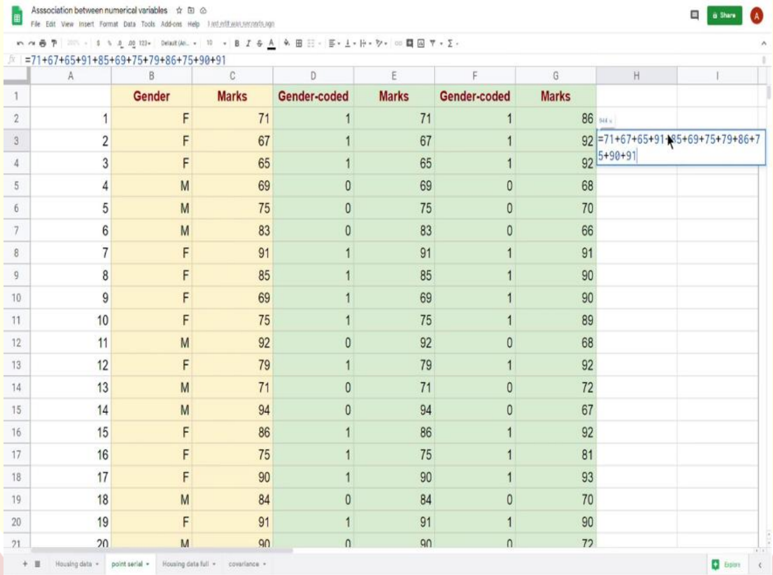
Step 2 Calculate the mean values of two groups: Let $\overset{\text{Male}}{\bar{Y}_0}$ and $\overset{\text{Female}}{\bar{Y}_1}$ be the mean values of groups with $Y = 0$, and $Y = 1$, respectively.

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.

n:- Total no. of obs.
n_0: # of obs. "0" a grp'
n_1 = # of obs "1".

The correlation coefficient

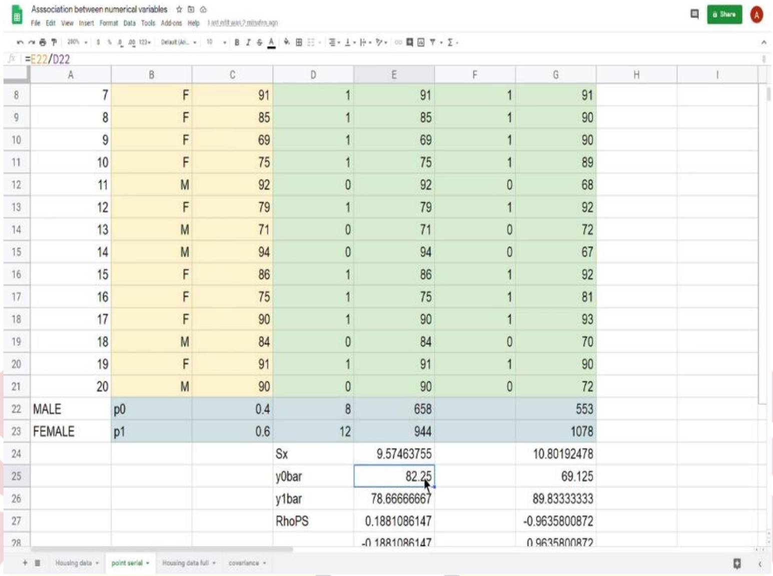
$$r_{pb} = \left(\frac{\bar{Y}_0 - \bar{Y}_1}{s_X}\right)\sqrt{p_0 p_1} \qquad \sqrt{\frac{n_0}{n-1}\cdot\frac{n_1}{n}}$$



Asssociation between numerical variables
File Edit View Insert Format Data Tools Add-ons Help
Share
A
Housing data
point serial
Housing data full
covariance
Explore

fx =71+67+65+91+85+69+75+79+86+75+90+91

	A	B	C	D	E	F	G	H	I
1		Gender	Marks	Gender-coded	Marks	Gender-coded	Marks		
2	1	F	71	1	71	1	86		
3	2	F	67	1	67	1	92	=71+67+65+91+85+69+75+79+86+75+90+91	
4	3	F	65	1	65	1	92		
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	M	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	1	91	1	90		
21	20	M	90	0	90	0	72		


Asssociation between numerical variables
File Edit View Insert Format Data Tools Add-ons Help
Share
Housing data
point serial
Housing data full
covariance

=E22/D22

	A	B	C	D	E	F	G	H	I
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	M	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	1	91	1	90		
21	20	M	90	0	90	0	72		
22	MALE	p0	0.4	8	658		553		
23	FEMALE	p1	0.6	12	944		1078		
24				Sx	9.57463755		10.80192478		
25				y0bar	82.25		69.125		
26				y1bar	78.66666667		89.83333333		
27				RhoPS	0.1881086147		-0.9635800872		
28					-0.1881086147		0.9635800872		

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⍰20⍰ 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.

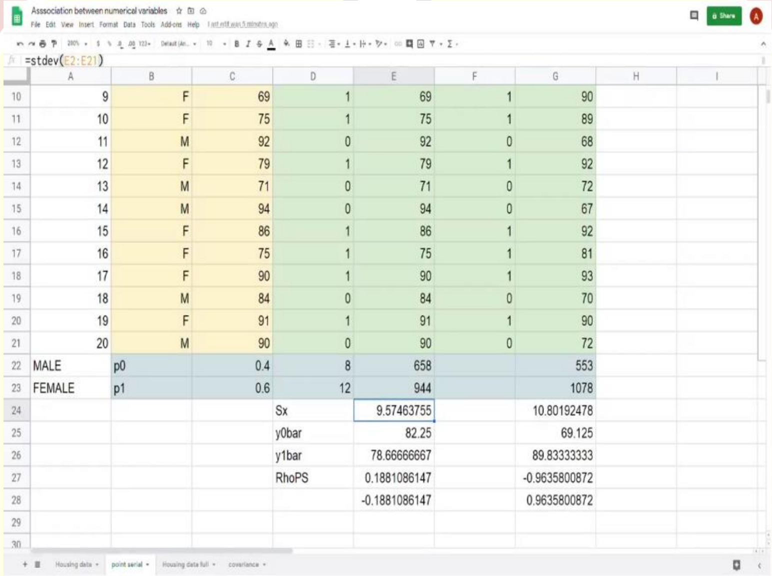
Similarly, 658 is the total marks obtained by all men students or male students put together. Now, what is the mean of this group? I have 12 students in this group, so the mean of this group which is going to be nothing but 944 ⍰12⍰which is equal to 78.66. So, $\overline{y_1}$ is giving me the mean of 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 y⍰0⍰⍰ and $\overline{y_1}$ which are the mean values of the group.

So, y⍰0⍰⍰ 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⍰, $p_1 = \frac{12}{20}$, 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.

So, S⍰X⍰ is the standard deviation of this variable. So, variable, entire variable which is $E2$ to $E21$. Then the point serial, point bi-serial correlation measure is defined as y⍰0⍰⍰− y⍰1⍰⍰⍰ S⍰X⍰⍰⍰⍰ divided by P_0P_1. Certain books use n⍰0⍰ and n_1, n_0 is the number of observations. So, p⍰0⍰ is basically n_0 by total let, if n is the total number of 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 $\frac{n_0}{n} = p_0$, p_1 is $\frac{n_1}{n}$. 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 $\frac{n_0}{n-1} \times \frac{n_1}{n}$. 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)



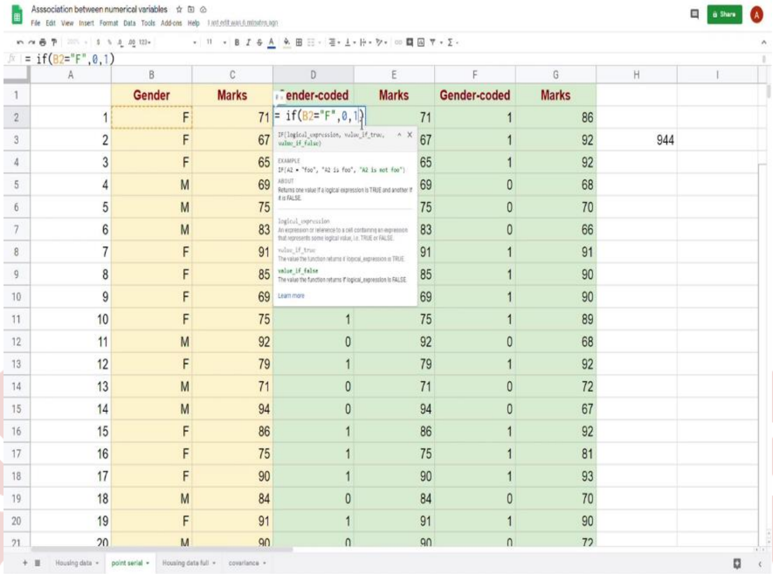
Asssociation between numerical variables

=stdev(E2:E21)

	A	B	C	D	E	F	G	H	I
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	M	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	1	91	1	90		
21	20	M	90	0	90	0	72		
22	MALE	p0	0.4	8	658		553		
23	FEMALE	p1	0.6	12	944		1078		
24				Sx	9.57463755		10.80192478		
25				y0bar	82.25		69.125		
26				y1bar	78.66666667		89.83333333		
27				RhoPS	0.1881086147		-0.9635800872		
28					-0.1881086147		0.9635800872		
29									
30									

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)


Asssociation between numerical variables
File Edit View Insert Format Data Tools Add-ons Help
Share
Housing data
point serial
Housing data full
covariance

`= if(B2="F",0,1)`

	A	B	C	D	E	F	G	H	I
1		Gender	Marks	ender-coded	Marks	Gender-coded	Marks		
2	1	F	71	= if(B2="F",0,1)	71	1	86		
3	2	F	67		67	1	92	944	
4	3	F	65		65	1	92		
5	4	M	69		69	0	68		
6	5	M	75		75	0	70		
7	6	M	83		83	0	66		
8	7	F	91		91	1	91		
9	8	F	85		85	1	90		
10	9	F	69		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	M	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	1	91	1	90		
21	20	M	90	0	90	0	72		

IF(logical_expression, value_if_true, value_if_false)

EXAMPLE
IF(A2 = "foo", "A2 is foo", "A2 is not foo")

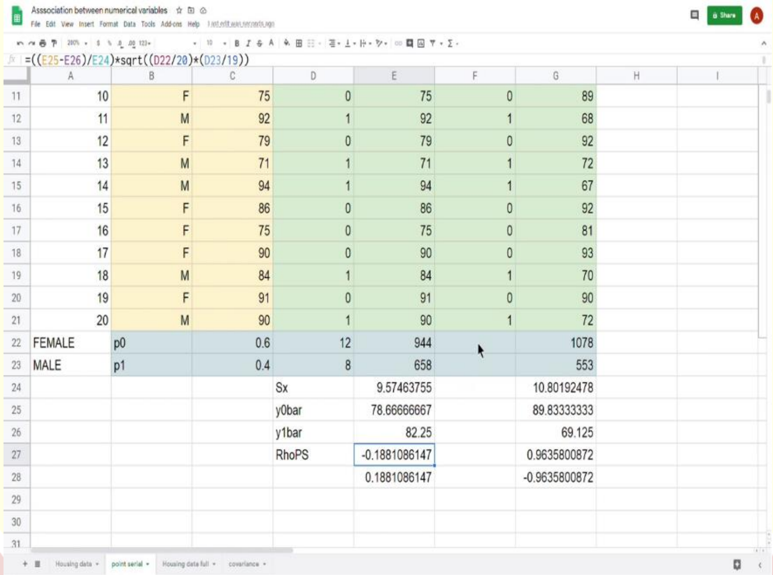
ABOUT
Returns one value if a logical expression is TRUE and another if it is FALSE.

logical_expression
An expression or reference to a cell containing an expression that represents some logical value, i.e. TRUE or FALSE.

value_if_true
The value the function returns if logical_expression is TRUE.

value_if_false
The value the function returns if logical_expression is FALSE.

Learn more


Asssociation between numerical variables
File Edit View Insert Format Data Tools Add-ons Help
Share
Housing data
point serial
Housing data full
covariance

`=((E25-E26)/E24)*sqrt((D22/20)*(D23/19))`

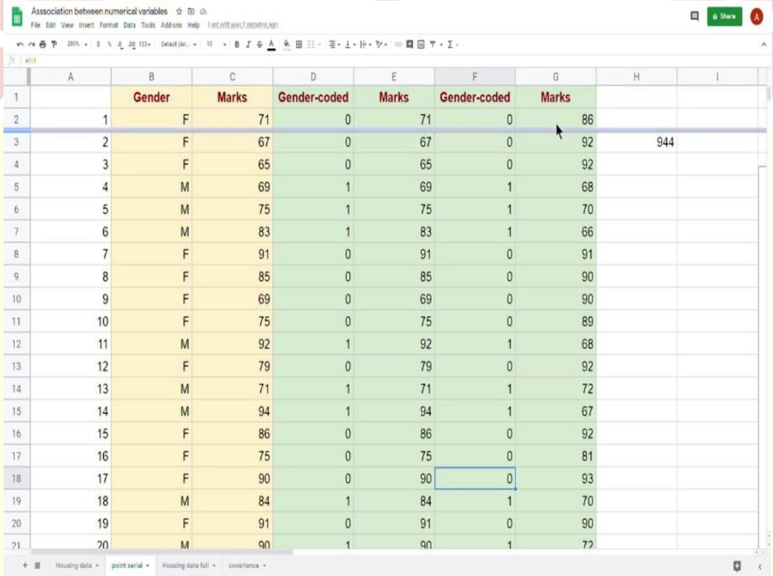
	A	B	C	D	E	F	G	H	I
11	10	F	75	0	75	0	89		
12	11	M	92	1	92	1	68		
13	12	F	79	0	79	0	92		
14	13	M	71	1	71	1	72		
15	14	M	94	1	94	1	67		
16	15	F	86	0	86	0	92		
17	16	F	75	0	75	0	81		
18	17	F	90	0	90	0	93		
19	18	M	84	1	84	1	70		
20	19	F	91	0	91	0	90		
21	20	M	90	1	90	1	72		
22	FEMALE	p0	0.6	12	944		1078		
23	MALE	p1	0.4	8	658		553		
24				Sx	9.57463755		10.80192478		
25				y0bar	78.66666667		89.83333333		
26				y1bar	82.25		69.125		
27				RhoPS	-0.1881086147		0.9635800872		
28					0.1881086147		-0.9635800872		
29									
30									
31									

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□1□□ 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.

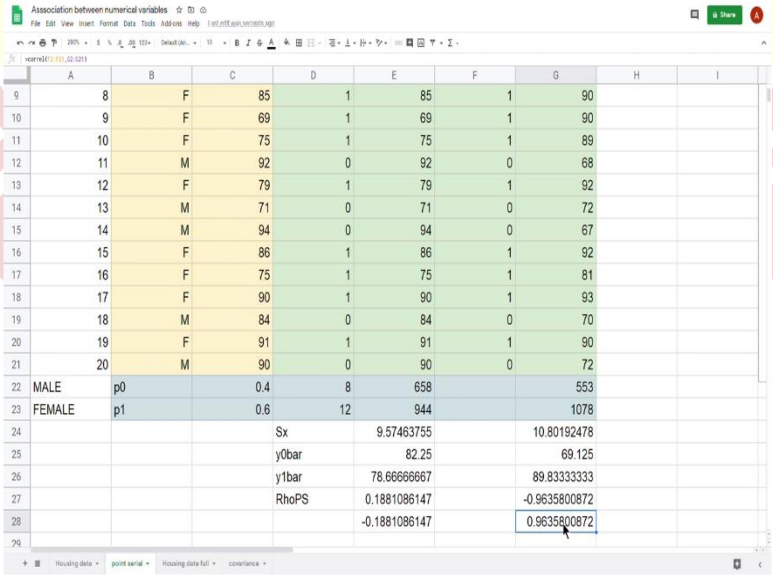
So, whether it is a very, so here I can, so basically in absolute terms this would lie between 0 and 1 and it is closer to 0, it says that in this dataset my gender really has no association with the marks. And that is sort of validated by my scatter plot also for my first example which said that irrespective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

(Refer Slide Time: 20:15)



Asssociation between numerical variables
File Edit View Insert Format Data Tools Add-ons Help
Share
Housing data
point serial
Housing data full
covariance

	A	B	C	D	E	F	G	H	I
1		Gender	Marks	Gender-coded	Marks	Gender-coded	Marks		
2	1	F	71	0	71	0	86		
3	2	F	67	0	67	0	92	944	
4	3	F	65	0	65	0	92		
5	4	M	69	1	69	1	68		
6	5	M	75	1	75	1	70		
7	6	M	83	1	83	1	66		
8	7	F	91	0	91	0	91		
9	8	F	85	0	85	0	90		
10	9	F	69	0	69	0	90		
11	10	F	75	0	75	0	89		
12	11	M	92	1	92	1	68		
13	12	F	79	0	79	0	92		
14	13	M	71	1	71	1	72		
15	14	M	94	1	94	1	67		
16	15	F	86	0	86	0	92		
17	16	F	75	0	75	0	81		
18	17	F	90	0	90	0	93		
19	18	M	84	1	84	1	70		
20	19	F	91	0	91	0	90		
21	20	M	90	1	90	1	72		



Asssociation between numerical variables
File Edit View Insert Format Data Tools Add-ons Help
Share
Housing data
point serial
Housing data full
covariance

=correl(F2:F21,G2:G21)

	A	B	C	D	E	F	G	H	I
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	M	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	1	91	1	90		
21	20	M	90	0	90	0	72		
22	MALE	p0	0.4	8	658		553		
23	FEMALE	p1	0.6	12	944		1078		
24				Sx	9.57463755		10.80192478		
25				y0bar	82.25		69.125		
26				y1bar	78.66666667		89.83333333		
27				RhoPS	0.1881086147		-0.9635800872		
28					-0.1881086147		0.9635800872		
29									

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.