June 2, 2021

WEEK 4

Association Between 2 Categorical Variables

LEARNING OBJECTIVES (for Week 4)

(1) Use of two way contingency tables to understand association between two categorical variables.

(2) Understand association between numerical variables through scatter plots ; compute and interpret correlation

(3) Understand relationship between a categorical and numerical variable

---

INTRODUCTION

- To understand the association between two categorical variables.
- Learn how to construct two-way contingency table
- Learn concept of relative row/column frequencies and how to use them to determine whether there is an association between the categorical variables.

EXAMPLE 1 (Nominal Variables)

Gender v/s Use of Smartphone

- A market research firm is interested in finding out whether ownership of a smartphone is associated with gender of a student. In other words, they want to find out whether more females owns a smartphone while compared to males, or whether owning a smartphone is independent of gender.

- To answer this question, a group of 100 college going children were surveyed about whether they owned a smartphone or not.

- The categorical variables in this example are:
  - → Gender : Male, Female (Nominal)
  - → Own a smartphone : Yes, No (Nominal)

Summarize Data :

→ We have the following summary statistics

(1) There are 44 female & 56 male students
(2) 76 students owned a smartphone, 24 did not own.
(3) 34 female students owned a smartphone, 42 male students owned a smartphone

→ The data given in the example can be organised using a two way table, referred to as contingency table.

	Own a smartphone		
Gender	No	Yes	Row Total
Female	10	34	44
Male	14	42	56
Col. Total	24	76	100

EXAMPLE 2 (Nominal & Ordinal Variable)

Income Vs Use of Smartphone

→ The categorical variables in this example are
- Income : Low, Medium, High (Ordinal)
- Own a smartphone : Yes, No (Nominal)

Contingency Table (summarizing data)

→ We have the foll. summary statistics :

(1) There are 20 high income, 66 medium income and 14 low income participants.

(2) 62 participants owned a smartphone, 38 did not own.

(3) 18 high income participants, 39 medium incomes participants and 5 low income participants owned a smartphone.

The contingency table corresponding to the data is given below:

Income Level	Own a smartphone		
	NO	YES	Row Total
HIGH	2	18	20
MEDIUM	27	39	66
LOW	9	5	14
Column Total	38	62	100

SECTION SUMMARY

- Organize bivariate categorical data into a two-way table : contingency table
- If data is ordinal, maintain order of the variable in the table.

June 3, 2021

RELATIVE FREQUENCY

ROW RELATIVE FREQUENCIES

→ What proportion of total participants own a smartphone ?

→ What proportion of female participants own a smartphone ?

Gender	Own a smartphone		
	NO	YES	Row Total
FEMALE	10	34	44
MALE	14	42	56
Column Total	24	76	100

* Row Relative Frequency : Divide each cell frequency in a row by its row total

Example 1 : GENDER Vs OWN A SMARTPHONE

Gender	Own a smartphone		
	NO	YES	Row Total
FEMALE	10/44	34/44	44
MALE	14/56	42/56	56
Column Total	24/100	76/100	100

Gender	Own a Smartphone		
	NO	YES	Row Total
FEMALE	22.73%	77.27%	44
MALE	25.00%	75.00%	56
Column Total	24.00%	76.00%	100

COLUMN RELATIVE FREQUENCIES

→ What proportion of total participants are females?

→ What proportion of smartphone owners are females?

* Column Relative Frequency : Divide each cell frequency in a column by its column total.

EXAMPLE : GENDER V/S OWN A SMARTPHONE

GENDER	OWN A SMARTPHONE		
	NO	YES	Row Total
FEMALE	10/24 = 41.67%	34/76 = 44.74%	44/100 = 44%
MALE	14/24 = 58.33%	42/76 = 55.26%	56/100 = 56%
Column Total	24	76	100

ASSOCIATION BETWEEN TWO VARIABLES

- What do we mean by stating two variables are associated?
  Knowing information about one variable provides information about the other variable.

→ To determine if two categorical variables are associated, we use the notion of relative row frequencies and relative column frequencies.

(1)→ If the row relative frequencies (the column relative frequencies) are the <u>same</u> for all rows (columns), then we say that two variables are <u>not associated</u> with each other.

(2)→ If the row relative frequencies (the column relative frequencies) are different for some rows (columns) then we say that the two variable are <u>associated</u> with each other.

EXAMPLE 1

Gender Vs Smartphone Ownership

Gender	Own a smartphone		
	NO	YES	Row Total
FEMALE	22.73%	77.27%	44
MALE	25.00%	75.00%	56
Column Total	24.00%	76.00%	100

' Row Frequency '

→ Now there we can see that there is not much difference in the relative frequencies of row 1, 2, 3. So in accordance to point (1), the two variables are not associated.

i.e, Gender & Smartphone Ownership are not associated.

EXAMPLE 2

Income Vs Smartphone Ownership

INCOME	OWNERSHIP		Row Total
(row frequency)	No	Yes	
High	10%	90%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100

→ Now, there we can see that the relative frequencies of all the three or four rows are very much different from each other. Hence, from point (2), we know that the two variables are associated.

i.e, Income & Smartphone Ownership are associated.

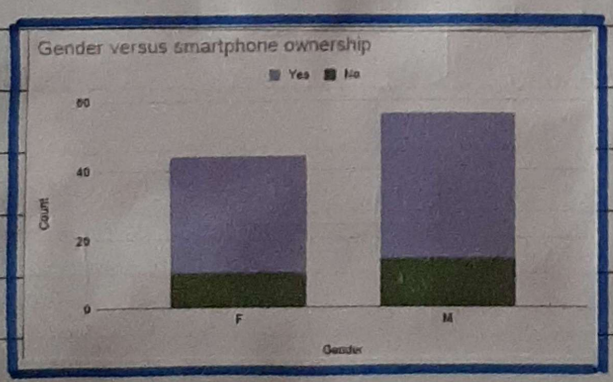
STACKED BAR CHART

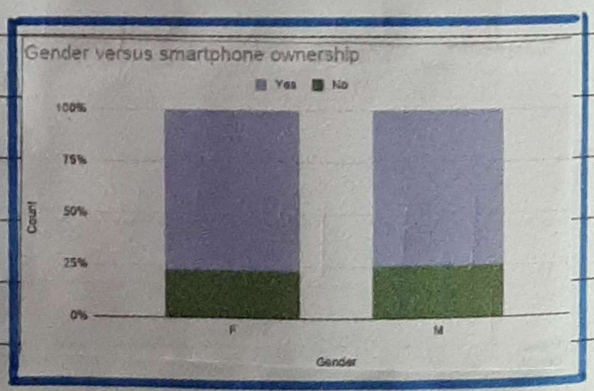
- Recall, a bar chart summarized the data for a categorical variable. It presented a graphical summary of the categorical variable under consideration with the length of the bars representing the frequency of occurence of a particular category.

- A STACKED BAR CHART represents the counts for a particular category. In addition, each bar is further broken down into smaller segments, with each segment representing the frequency of that particular category within the segment. A stacked bar chart is also referred to as a segmented bar chart.

EXAMPLE 1

	Own a smartphone		
Gender	No	Yes	Row total
Female	22.73%	77.27%	44
Male	25.00%	75.00%	56
Column total	24.00%	76.00%	100

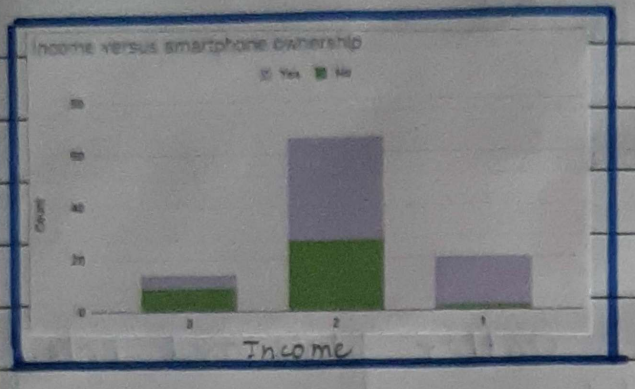


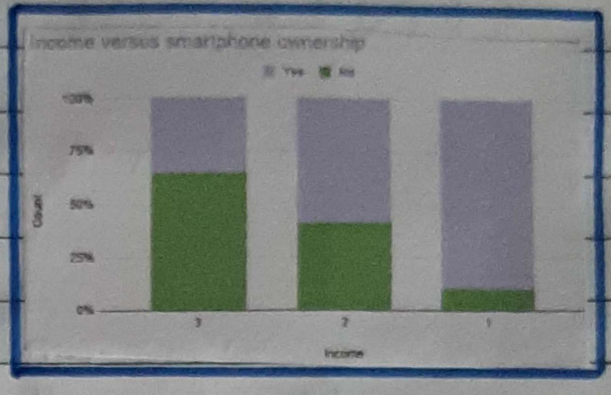


A 100% stacked bar chart is useful to part-to-whole relationships

EXAMPLE 2

	Owns a smartphone		
Income level	No	Yes	Row Total
High	10.00%	90.00%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100





Frequency, Stacked Bar Chart

100% stacked bar Chart

Date
June 4, 2021

classmate
Date
Page

Association Between Two Numerical Var.

INTRODUCTION

- To understand the association between two numerical variables.
- Learn how to construct scatter plots and interpret association in scatter plots
- Summarize association with a line
- Correlation matrix

SCATTER PLOT

- We use a scatterplot to look for association between numerical variables

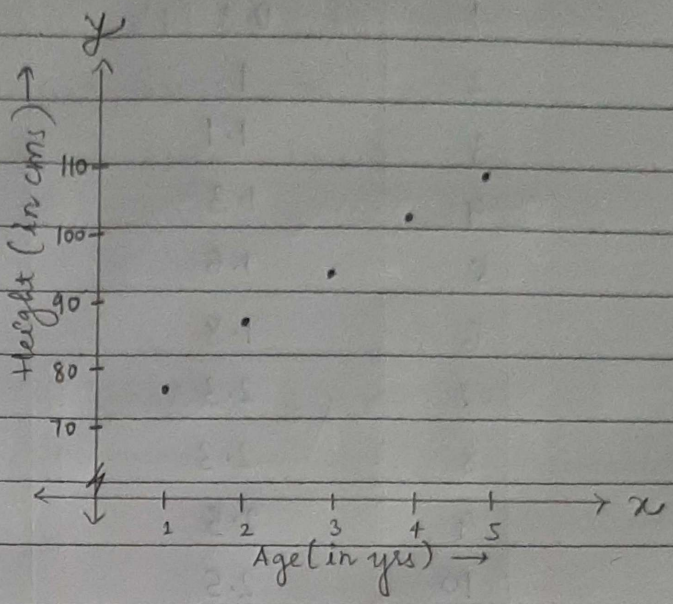
Definition: A scatter plot is a graph that displays pairs of values as points on a two-dimensional plane.

→ To decide which variable to put on the x-axis and which to put on y-axis, display the variable you would like to explain along the y-axis (referred as

response variable) and the variable which explains on x-axis (referred as explanatory variable).

EXAMPLE 1

AGE (in yrs)	HEIGHT (cms)
1	75
2	85
3	94
4	101
5	108



EXAMPLE 2

A real estate agent collected the prices of different size of homes. He wanted to see what was the relationship between the price of a home and size of home. In particular, he wanted to know if the prices of homes increased linearly with the size or in any other way?

To answer this question, he collected data on 15 homes. The data he recorded was

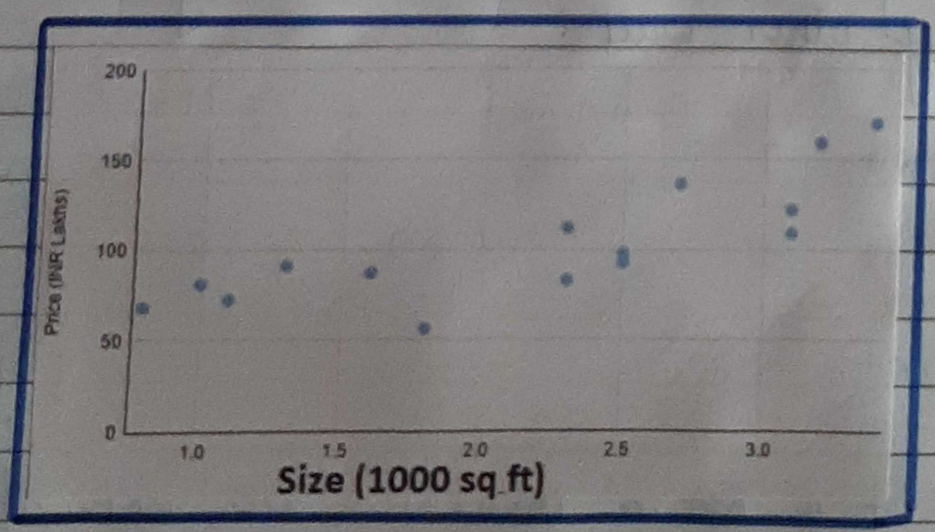
(1) Size of a home measured in 1000 of square feet

(2) Price of a home measured in lakh of rupees.

HOUSING DATA

S.no	Size (1000 sq ft)	Price (INR Lakhs)
1	0.8	68
2	1	81
3	1.1	72
4	1.3	91
5	1.6	87
6	1.8	56
7	2.3	83
8	2.3	112
9	2.5	93
10	2.5	98
11	2.7	136
12	3.1	109
13	3.1	122
14	3.2	159
15	3.4	170

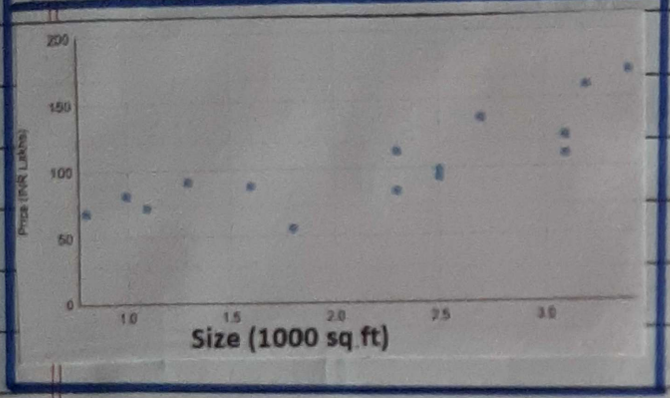
SCATTER PLOT



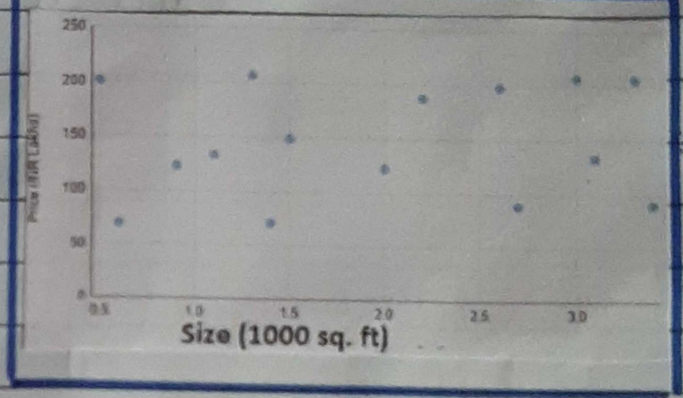
VISUAL TEST FOR ASSOCIATION

- Do we see a pattern in the scatter plot?
- In other words, if I know about the x-value, can I use it to say something about the y-value or guess y-value?

DATASET 1 (Scatter Plot)



DATASET 2 (Scatter Plot)



→ Here we can see that, Dataset 1 follows some kind of pattern but there is no pattern being followed in Dataset 2.

DESCRIBING ASSOCIATION

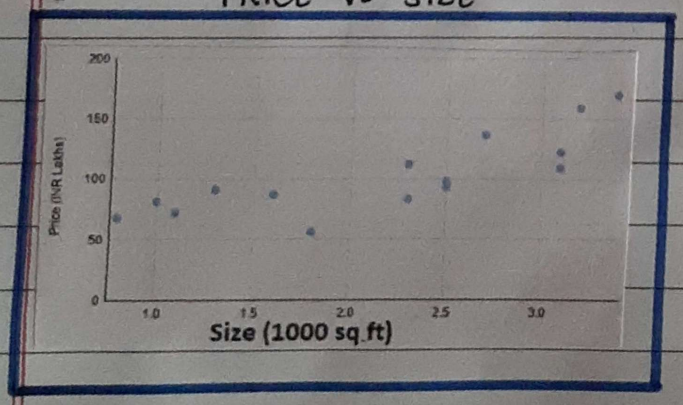
When describing association between variables in a scatter plot, there are four key questions that need to be answered.

1. DIRECTION : Does the pattern trend up, down, or both?
2. CURVATURE : Does the pattern appear to be linear or does it curve?
3. VARIATION : Are the points tightly clustered along the pattern?
4. OUTLIERS : Did you find something unexpected?

1. DIRECTION

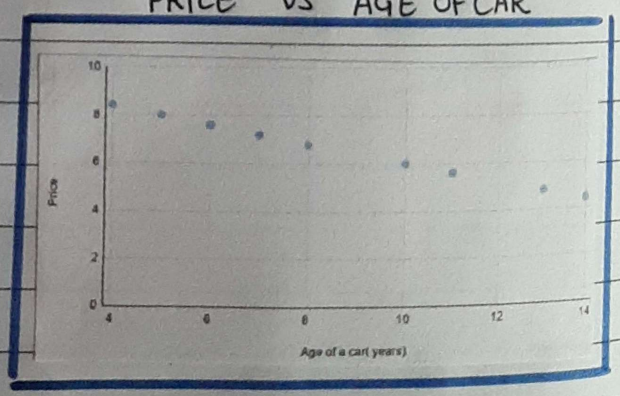
Does the pattern trend up, down or both?

PRICE VS SIZE



UP

PRICE VS AGE OF CAR

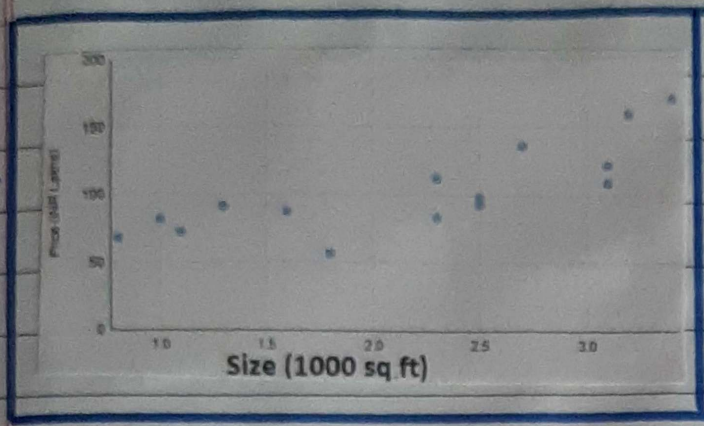


DOWN

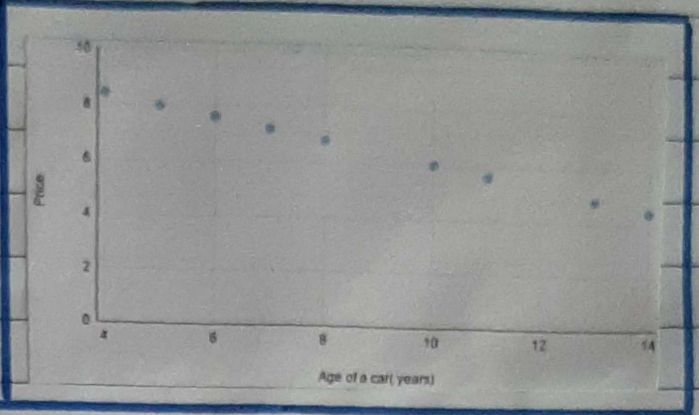
2. CURVATURE

Does the pattern appear to be linear or does it curve?

LINEAR

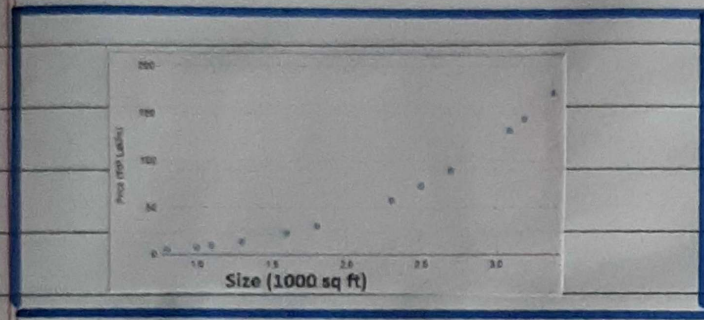


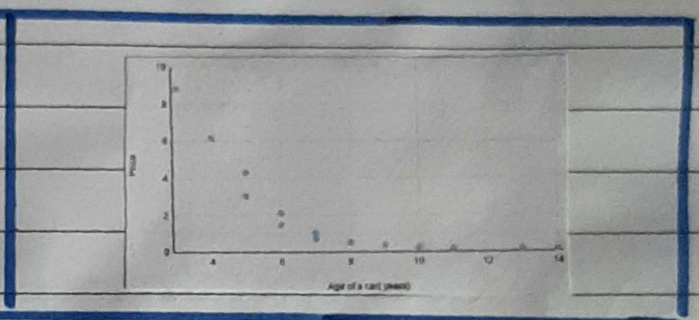
PRICE VS SIZE



PRICE VS AGE OF CAR

CURVE

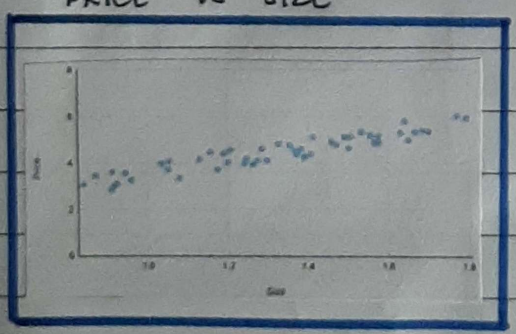




3. VARIATION

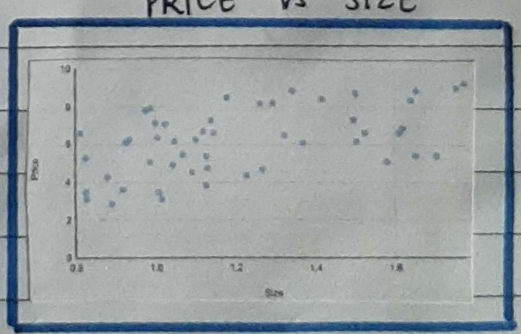
Are the points tightly clustered along the pattern?

PRICE VS SIZE



TIGHTLY CLUSTERED

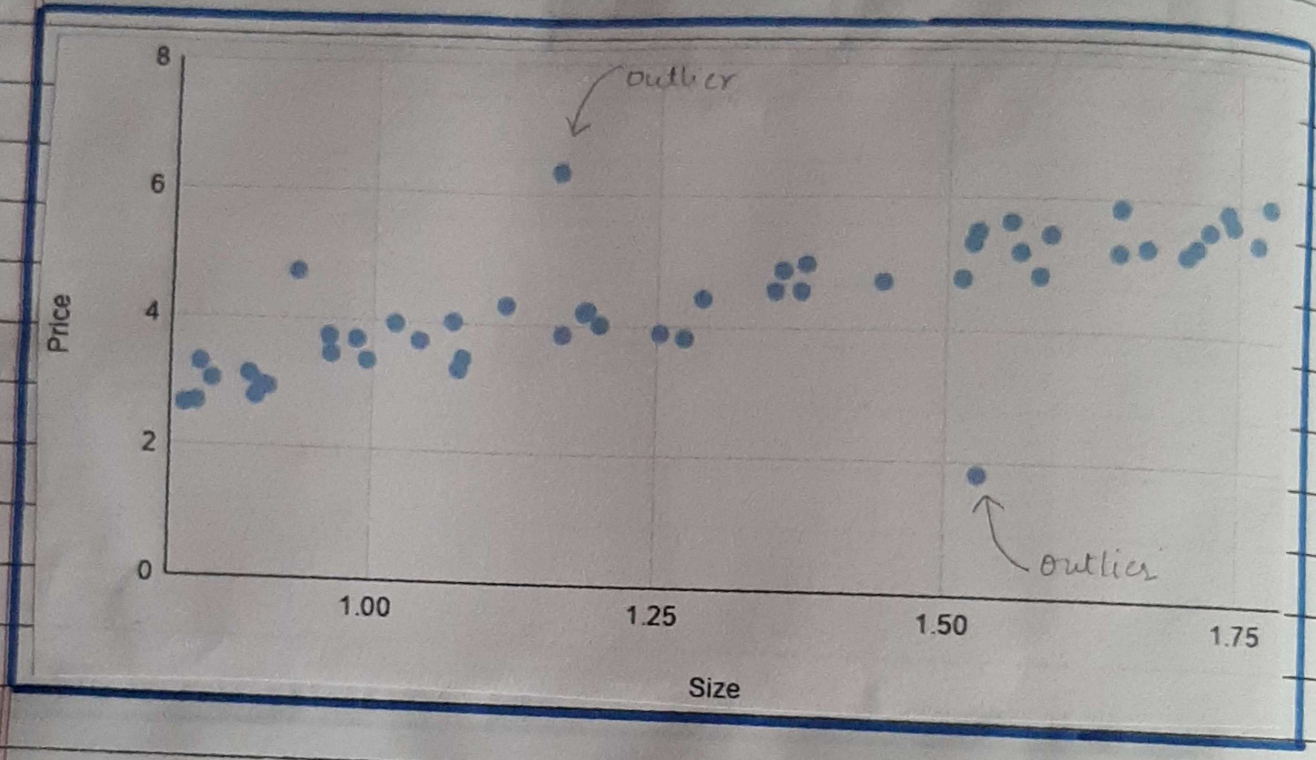
PRICE VS SIZE



VARIABLE

4. OUTLIERS

Did you find something unexpected ?



June 5, 2021

MEASURES OF ASSOCIATION

How do we measure the strength of association between two variables?

1. Covariance
2. Correlation

1. COVARIANCE

Covariance quantifies the strength of the linear association between two numerical variables.

EXAMPLE 1

Recall, the association between age and height of a person.

Age (in yrs) x	Height (in cms) y	Deviation of x $(x_i - \bar{x})$	Dev. of y $(y_i - \bar{y})$
1	75	-2	-17.6
2	85	-1	-7.6
3	94	0	1.4
4	101	1	8.4
5	108	2	15.4

$\bar{x} = 3$ $\bar{y} = 92.6$

$(x_i - \bar{x})(y_i - \bar{y})$
35.2
7.6
0
8.4
30.8

~~Pop variance, $\sigma^2 = \frac{82}{5} = 16.4$~~

~~Sam. variance, $s^2 = \frac{82}{4} = 20.5$~~

EXAMPLE 2 Variables : Age of a car & price of a car

Age (in yrs) x	Price (INR lakhs) Y	Dev. of x $(x_i - \bar{x})$	Dev of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	-2	2	-4
2	5	-1	1	-1
3	4	0	0	0
4	3	1	-1	-1
5	2	2	-2	-4
$\bar{x} = 3$	$\bar{y} = 4$	~~$s_x^2 = 10/5 = 2$~~	~~$s_y^2 = 10/4 = 2.5$~~	

Key Observation :

- When large (small) values of x tend to be associated with large (small) values of y - the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also be tend to be same.

- When large (small) values of x tend to be associated with small (large) values of y - the signs of deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be different.

X	Y	Signs of Dev.
Large	Large	Same
Small	Small	Same
Large	Small	Diff.
Small	Large	Diff.

Definition: Let x_i denote the i^{th} observation of variable x, and y_i denote the i^{th} observation of variable y. Let (x_i, y_i) be the i^{th} paired observation of a population (sample) dataset having $N(n)$ observations. The Covariance between the variables x and y is given by

→ Population Covariance : $$Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{N}$$

→ Sample Covariance : $$Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Units of Covariance:

→ The size of covariance, however, is difficult to interpret because the covariance has units.

→ The units of the covariance are those of x-variable times those of the y-variable.

Example 1 : Population variance $= \frac{-17.6 + (-7.6) + 1.4 + 8.4 + 15.4}{5}$

$= 82 \div 5 = 16.4$

Sample variance $= \frac{82}{4} = 20.5$

Example 2 : Population variance $= \frac{-4 + (-1) + 0 + (-1) + (-4)}{5} = \frac{-10}{5}$

$= -2$

Sample Variance $= \frac{-10}{4} = -2.5$

2. CORRELATION

- A more easily interpretted measure of linear association between two numerical variables is correlation.

- It is derived from covariance.

- To find the correlation between two numerical variables x and y divide the covariance between x and y by the product of the std. deviations of x and y. The pearson correlation coefficient, r, between x & y is given by

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{cov(x,y)}{s_x s_y}$$

NOTE :
- → The units of the standard deviations cancel out the units of covariance.

- → It can be shown that the correlation measure always lies between -1 and $+1$.

EXAMPLE 1

Age x	Height Y	sq. Devn of x $(x_i - \bar{x})^2$	sq. Devn of y $(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	4	309.76	35.2
2	85	1	57.76	7.6
3	94	0	1.96	0
4	101	1	70.56	8.4
5	108	4	237.16	30.8
		10	677.2	82

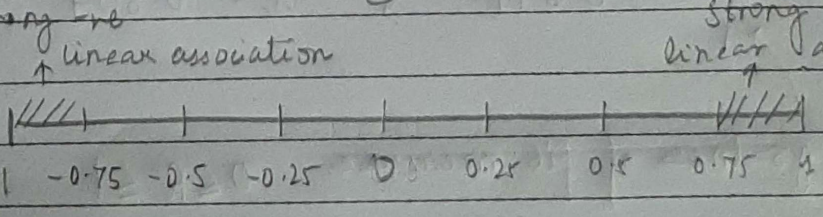
→ $S_x = 1.58$, $S_y = 13.01$

→ $r = \frac{82}{\sqrt{10 \times 677.2}}$ or $\frac{20.5}{1.58 \times 13.01} = 0.9964$

Covariance → if $cov(x, y) \rightarrow +ve$
It means both are going up

if $cov(x, y) \rightarrow -ve$
One is going up & other down.

Correlation :



FITTING A LINE

Learning Objectives

→ Summarize the linear association between two variables using the equation of a line.

→ Understand the significance of R^2

SUMMARIZING THE ASSOCIATION WITH A LINE

- The strength of linear association between the variables was measured using the measures of covariance & correlation.
- The linear association can be described using the equation of a line.

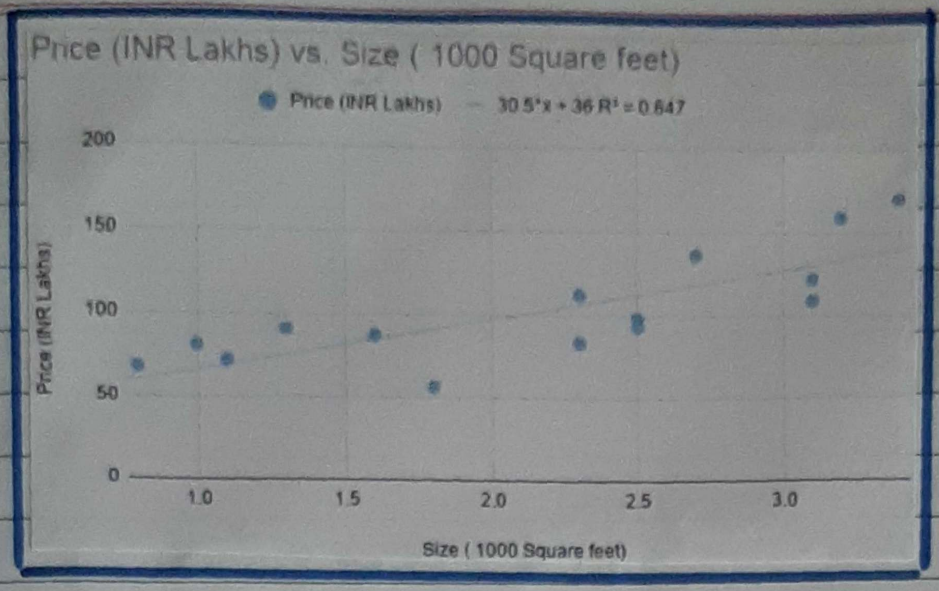
EXAMPLE 1 SIZE Versus PRICE OF HOMES : Equation

$y = mx + c$

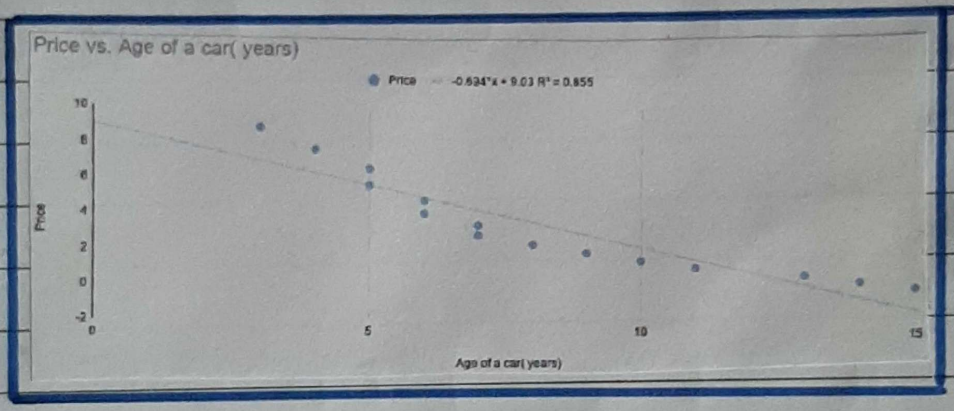
Equation of the line : Price = 30.5 X Size + 36

$R^2 = 0.647$; $r = 0.804$

goodness of fit measure $0 \leq R^2 \leq 1$



EXAMPLE 2 AGE VS PRICE OF CARS : Equation



Equation of Line : Price = $-0.694 \times$ Age $+ 9.03$

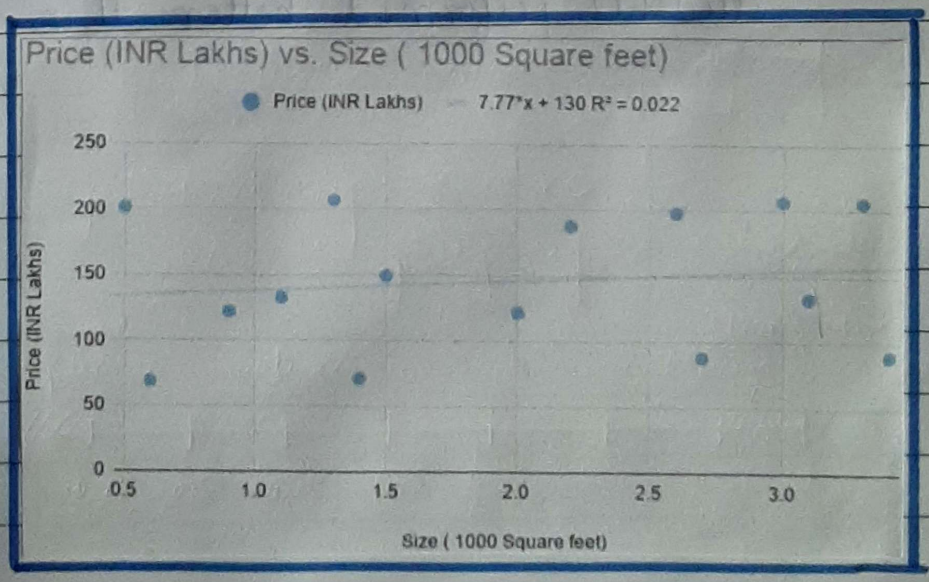
$R^2 = 0.855$; $r = -0.9247$

EXAMPLE 3 :

Eqn of Line :

Price = $7.77 \times$ Size $+ 130$

$R^2 = 0.022$; $r = 0.149$



June 6, 2021

Association Between Categorical & Numerical Variables

INTRODUCTION

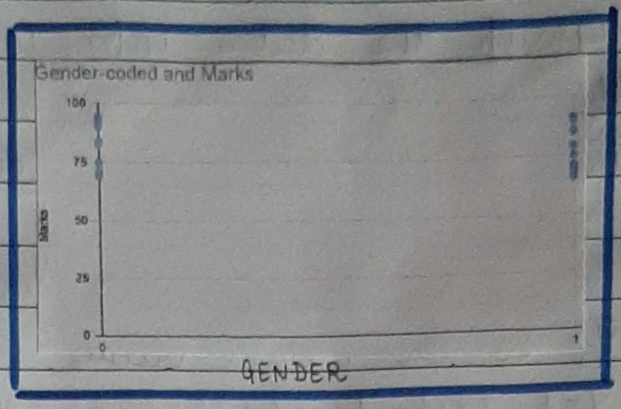
- Understand the association between a categorical variable and numerical variable.
- Assume the categorical variable has two categories (dichotomous)

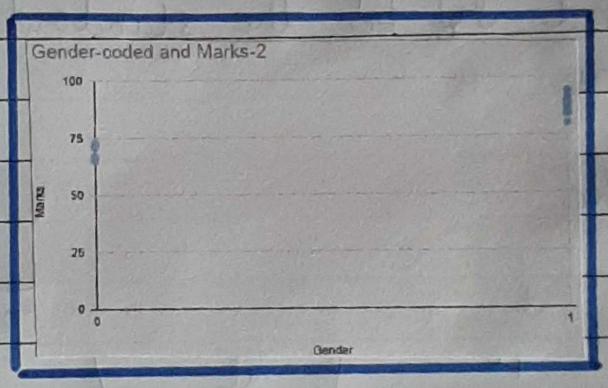
EXAMPLE 1 : GENDER VERSUS MARKS

A teacher was interested in knowing if female students performed better than male students in her class. She collected data from 20 students and the marks they obtained on 100 in the subject.

Sno.	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

SCATTER PLOT





Another Dataset

POINT BI-SERIAL CORRELATION COEFFICIENT

→ Let X be a numerical variable and Y be a categorical variable with 2 categories (a dichotomous var.)

→ 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 $\bar{Y}_0$ and $\bar{Y}_1$ 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.

The correlation coefficient

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

$p_0 = \frac{\text{no. of obs. in 0}}{\text{total obs.}}$

S_x ← stand. dev