Week 4

Learning objectives

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

I To understand the association between two categorical variables.

I Learn how to construct two-way contingency table.

I 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: Gender versus use of smartphone

I 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 own a smartphone while compared to males, or whether owning a smartphone is independent of gender.

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

I The categorical variables in this example are

I Gender: Male, Female (2 categories)- Nominal variable

I Own a smartphone: Yes, No (2 categories)- Nominal variable

Example 1: Gender versus use of smartphone-summarize data

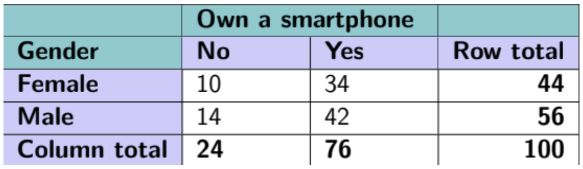
I We have the following summary statistics

1. There are 44 female and 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.

I The data given in the example can be organized using a two-way table, referred to as a contingency table.



	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	10	34	**44**
Male	14	42	**56**
Column total	**24**	**76**	**100**

Contingency table using google sheets

Step 1 Choose the columns of the variables for which you seek anassociation.

Step 2 Go to Data-click on Pivot table option

Step 3 Click on create option in the pivot table-it will open the pivot table editor:

3.1 Under the Rows tab, click on the first categorical variable.

3.2 Under the columns tab, click on the second categoricalvariable.

3.3 Under the values tab, click on either of the variables and then click on the COUNTA tab under "summarize by" tab.

Example 2: Income versus use of smartphone

I A market research firm is interested in finding out whetherownership of a smartphone is associated with income of an individual. In other words, they want to find out whether income is associated with ownership of a smartphone.

I To answer this question, a group of 100 randomly picked individuals were surveyed about whether they owned a smart phone or not.

I The categorical variables in this example are

I Income: Low, Medium, High (3 categories) -Ordinal variable

I Own a smartphone: Yes, No (2 categories) - Nominal variable

Example 2: Contingency table

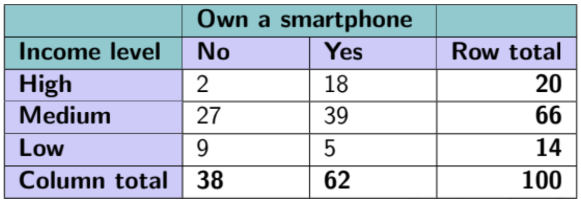
I We have the following summary statistics

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 owned a smartphone, 39 Medium income participants owned a smartphone, and 5 Low income participants owned a smartphone.

I The contingency table corresponding to the data is given below.



	Own a smartphone		
Income level	No	Yes	Row total
High	2	18	20
Medium	27	39	66
Low	9	5	14
Column total	38	62	100

<u>Section summary</u>

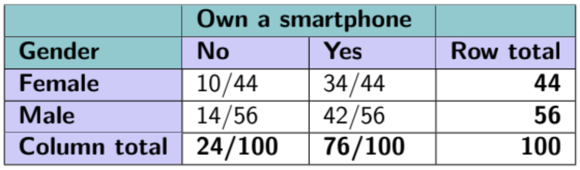
I Organize bivariate categorical data into a two-way table-contingency table.

I If data is ordinal, maintain order of the variable in the table.

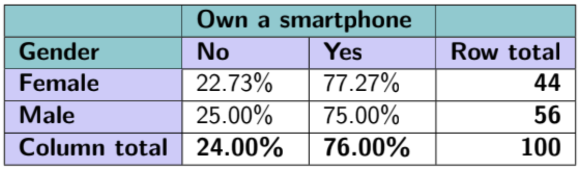
Row relative frequency: Divide each cell frequency in a row by its row total.

EXAMPLES

1.

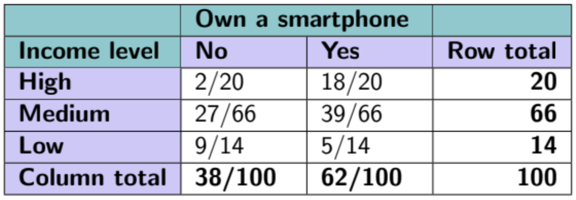


	Own a smartphone		
Gender	No	Yes	Row total
Female	10/44	34/44	44
Male	14/56	42/56	56
Column total	24/100	76/100	100

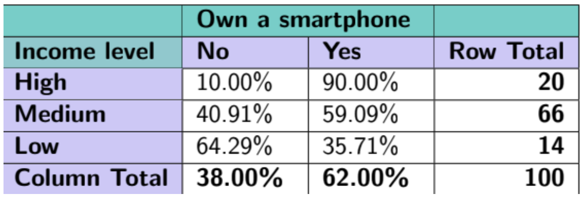


	Own a smartphone		
Gender	No	Yes	Row total
Female	22.73%	77.27%	44
Male	25.00%	75.00%	56
Column total	24.00%	76.00%	100

2.

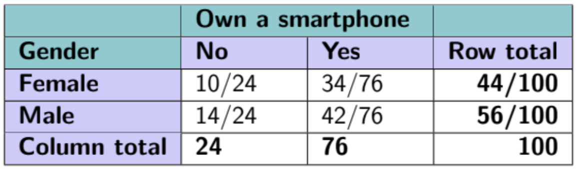


	Own a smartphone		
Income level	No	Yes	Row total
High	2/20	18/20	20
Medium	27/66	39/66	66
Low	9/14	5/14	14
Column total	38/100	62/100	100

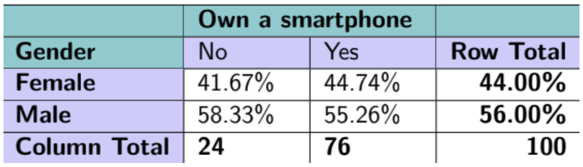


	Own a smartphone		
Income level	No	Yes	Row Total
High	10.00%	90.00%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100

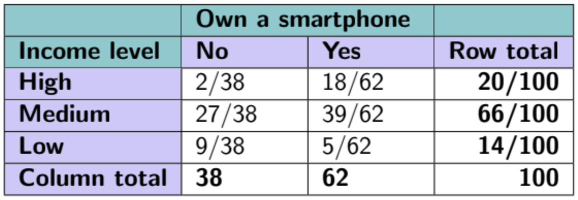
Column relative frequency: Divide each cell frequency in a column by its column total.



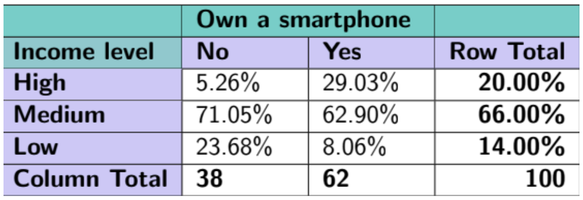
	Own a smartphone		
Gender	No	Yes	Row total
Female	10/24	34/76	44/100
Male	14/24	42/76	56/100
Column total	24	76	100



	Own a smartphone		
Gender	No	Yes	Row Total
Female	41.67%	44.74%	44.00%
Male	58.33%	55.26%	56.00%
Column Total	24	76	100



	Own a smartphone		
Income level	No	Yes	Row total
High	2/38	18/62	**20/100**
Medium	27/38	39/62	**66/100**
Low	9/38	5/62	**14/100**
Column total	**38**	**62**	**100**



	Own a smartphone		
Income level	No	Yes	Row Total
High	5.26%	29.03%	**20.00%**
Medium	71.05%	62.90%	**66.00%**
Low	23.68%	8.06%	**14.00%**
Column Total	**38**	**62**	**100**

<u>Association between two variables</u>

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

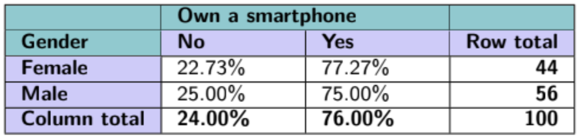
I To determine if two categorical variables are associated, we use the notion of relative row frequencies and relative column frequencies described earlier.

I If the row relative frequencies (the column relative frequencies) are the **same** for all rows (columns) then we say that the two variables are not associated with each other.

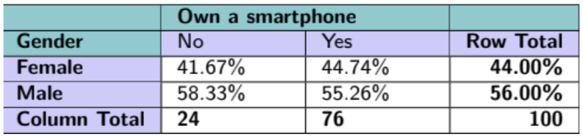
I If the row relative frequencies (the column relative frequencies) are **different** for some rows (some columns) then we say that the two variables are associated with each other.

Example 1: Association between two variables

I If the row relative frequencies (the column relative frequencies) are the same for all rows (columns) then we say that the two variables are not associated with each other.



	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**

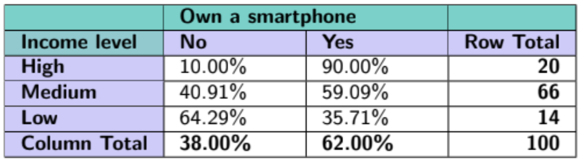


	Own a smartphone		
Gender	No	Yes	Row Total
Female	41.67%	44.74%	**44.00%**
Male	58.33%	55.26%	**56.00%**
Column Total	**24**	**76**	**100**

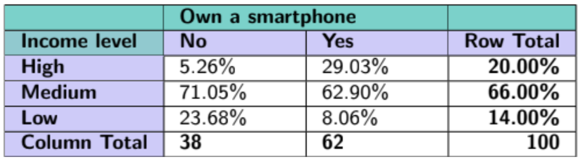
Gender and smartphone ownership are not associated

Example 2: Association between two variables

I If the row relative frequencies (the column relative frequencies) are different for some rows (some columns) then we say that the two variables are associated with each other.



	Own 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**



	Own a smartphone		
Income level	No	Yes	Row Total
High	5.26%	29.03%	**20.00%**
Medium	71.05%	62.90%	**66.00%**
Low	23.68%	8.06%	**14.00%**
Column Total	**38**	**62**	**100**

Income and smartphone ownership are associated

<u>Stacked bar chart</u>

I **bar chart** presented a graphical summary of the categorical variable under consideration, with the length of the bars representing the frequency of occurrence of a particular category.

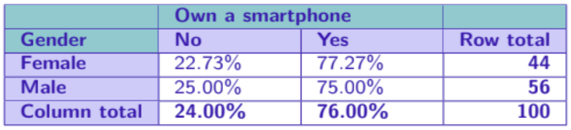
I A **stacked bar chart** represents the counts for a particular category. In addition, each bar is further broken down into smaller segments, with each segment representing the frequency of that particular category within the segment. A stacked bar chart is also referred to as a segmented bar chart.

Stacked bar chart using google sheets

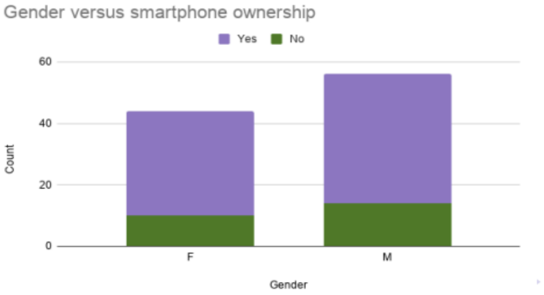
Step 1: Select the data you want to include in the contingency table.

Step 2: Click Insert - chart- choose stacked bar option

Example 1: Stacked bar chart

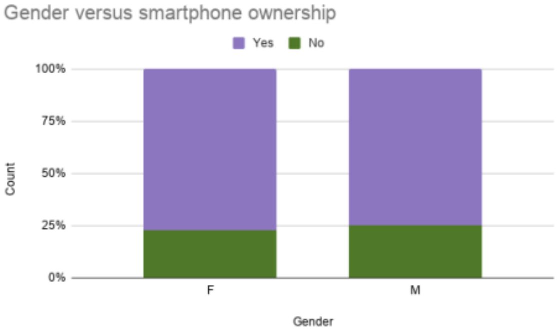


	Own a smartphone		
Gender	No	Yes	Row total
Female	22.73%	77.27%	44
Male	25.00%	75.00%	56
Column total	24.00%	76.00%	100

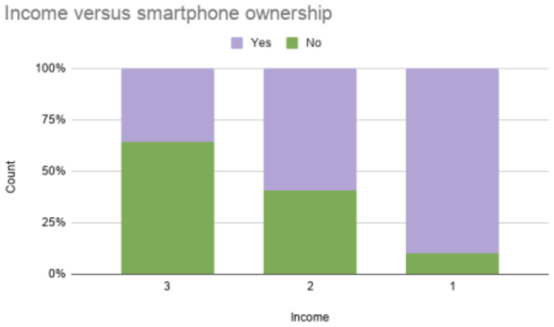


Example 1: 100% Stacked bar chart

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



Example 2: Stacked bar chart



Section summary

I Concept of relative frequency: row relative frequency and column relative frequency.

I Understand whether two categorical variables are associatedusing the concept of relative frequencies.

I Graphical summary of association using stacked bar chart.

Introduction

I To understand the association between two numerical variables.

I Learn how to construct scatter plots and interpret association in scatter plots.

I Summarize association with a line.

I Correlation matrix.

Scatter plot (to look for association between numerical variables**)**

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

I To decide which variable to put on the x-axis and which to put on the 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 : Prices of homes

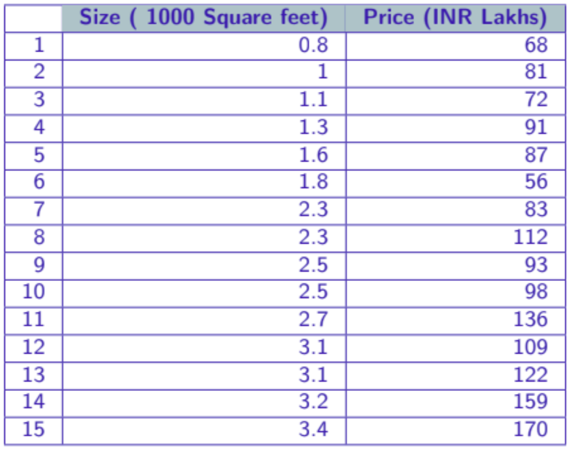
A real estate agent collected the prices of different sizes of homes. He wanted to see what was the relationship between the price of a home and size of a home. In particular, he wanted to know if the prices of homes increased linearly with the size or in any other way?

To answer the 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



	Size (1000 Square feet)	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 using google sheets

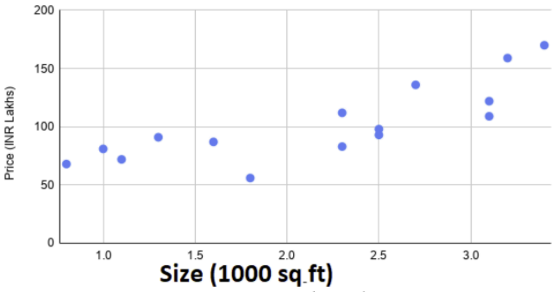
Step 1: Highlight data you want to plot

Step 2: Insert - chart- choose scatter chart

Step 3: Under X−axis tab, choose your explanatory variable.

Step 4: Under series tab, the response variable.

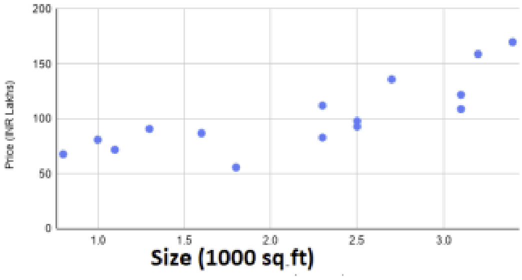
Step 5: Label the title of the chart, axes appropriately.



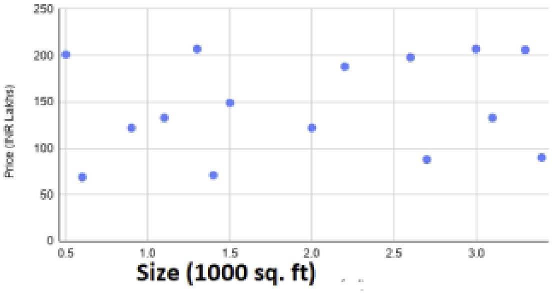
Visual test for association

I Do we see a pattern in the scatter plot?

I 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?



‘



Section summary

1. scatter plot , Visual test for association

2. Notion of explanatory variable and response variable.

Describing association

When describing association between varaibles 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

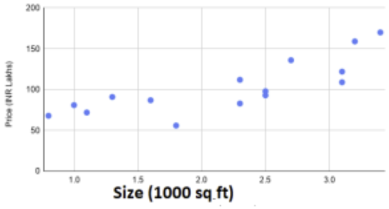
curve?

3. **Variation**: Are the points tightly clustered along the pattern?

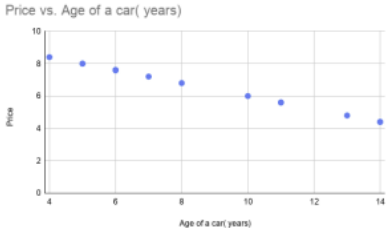
4. **Outliers**: Did you find something unexpected?

Describing association: Direction

Does the pattern trend up, down, or both?

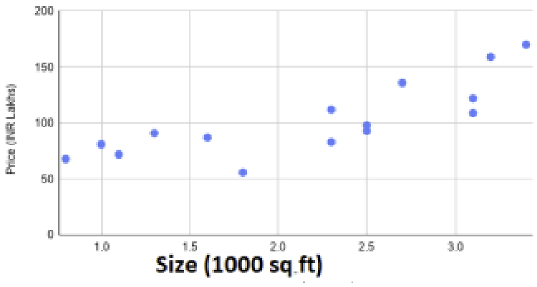


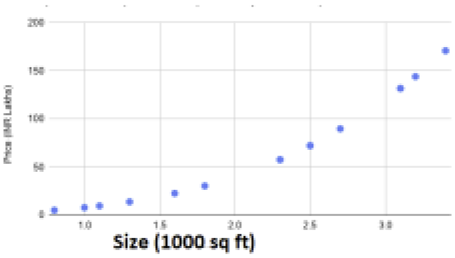
Up

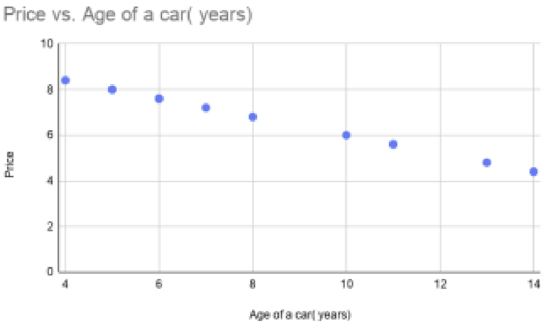


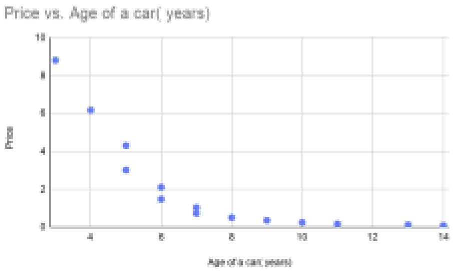
Down

Describing association: Curvature



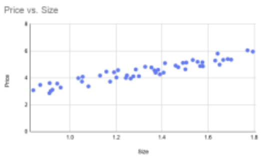




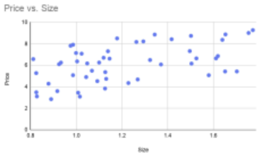


Describing association: Variation

Are the points tightly clustered along the pattern?



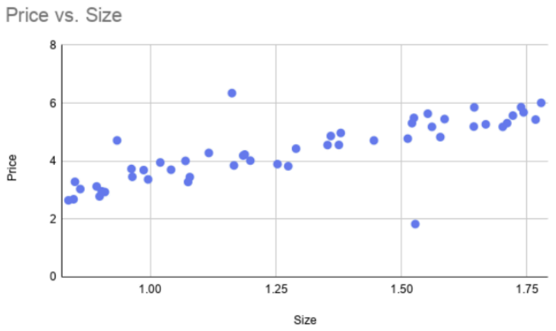
Tightly clustered



Variable

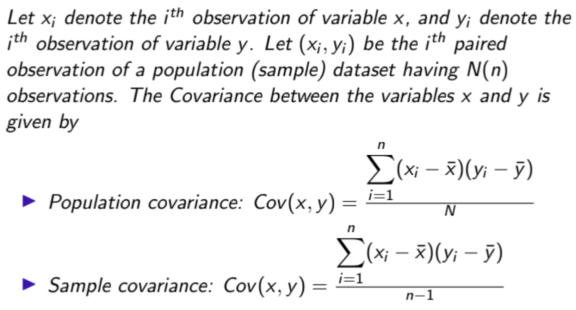
Describing association: Outliers

Did you find something unexpected?



Measure the strength of association between two variables – Covariance and Correlation

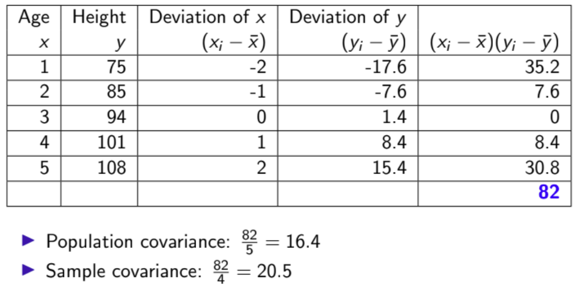
Covariance quantifies the strength of the linear association between two numerical variables.



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

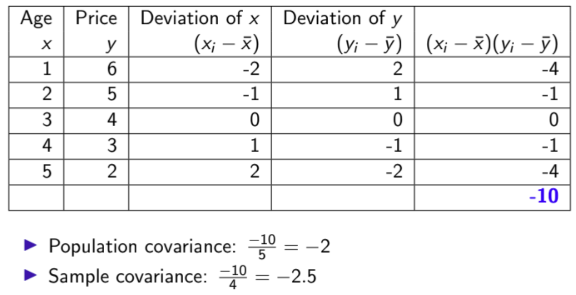
Ex-1.the association between age and height of a person.



Age	Height	Deviation of x	Deviation of y	
x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	-2	-17.6	35.2
2	85	-1	-7.6	7.6
3	94	0	1.4	0
4	101	1	8.4	8.4
5	108	2	15.4	30.8
				82

- Population covariance: $\frac{82}{5} = 16.4$
- Sample covariance: $\frac{82}{4} = 20.5$

2. Age of a car and price of a car



Age	Price	Deviation of x	Deviation of y	
x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
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
				-10

- Population covariance: $\frac{-10}{5} = -2$
- Sample covariance: $\frac{-10}{4} = -2.5$

Key observation

I When large (small) values of x tend to be associated with large (small) values of y- the signs of the deviations, (xi − x̄) and (yi − ȳ) will also tend to be **same.**

I When large (small) values of x tend to be associated with small (large) values of y- the signs of the deviations, (xi − x̄) and (yi − ȳ) will also tend to be **different.**

Units of Covariance

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

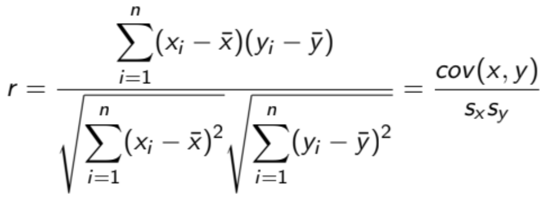
I The units of the covariance are those of the x-variable times those of the y-variable**.**

Correlation

I A more easily intepreted measure of linear association between two numerical variables is correlation

I It is derived from covariance.

I To find the correlation between two numerical variables x and y divide the covariance between x and y by the product of the standard deviations of x and y. The Pearson correlation coefficient, r, between x and y is given by



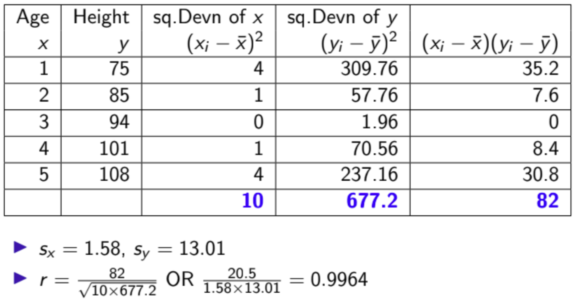
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{cov(x,y)}{s_x s_y}$$

Remark

1.The units of the standard deviations cancel out the units of covariance

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

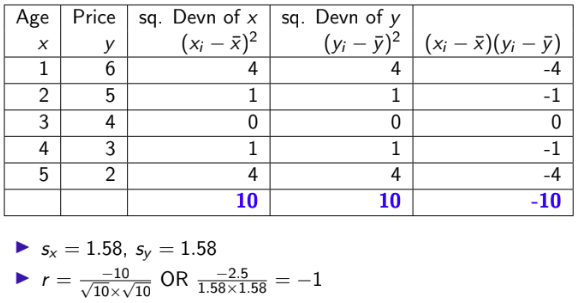
Correlation: Example 1



Age	Height	sq.Devn of x	sq.Devn of y	
x	y	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
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 $\frac{20.5}{1.58 \times 13.01} = 0.9964$

Correlation: Example 2



Age	Price	sq. Devn of x	sq. Devn of y	
x	y	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	4	4	-4
2	5	1	1	-1
3	4	0	0	0
4	3	1	1	-1
5	2	4	4	-4
		10	10	-10

- $s_x = 1.58,\ s_y = 1.58$
- $r = \frac{-10}{\sqrt{10} \times \sqrt{10}}$ OR $\frac{-2.5}{1.58 \times 1.58} = -1$

Correlation using google sheets

Step 1 The function CORREL(series1, series2) will return the value of correlation.

For example: If the data corresponding to x-variable (series1) is in cell A2:A6 and data corresponding to y-variable (series2) is in cells B2:B6; then CORREL(A2:A6,B2:B6) returns the value of the Pearson Correlation coefficient.

Summarizing the association with a line

I The strength of linear association between the variables was measured using the measures of Covariance and Correlation.

I The linear association can be described using the equation of a line.

Equation of line using google sheets

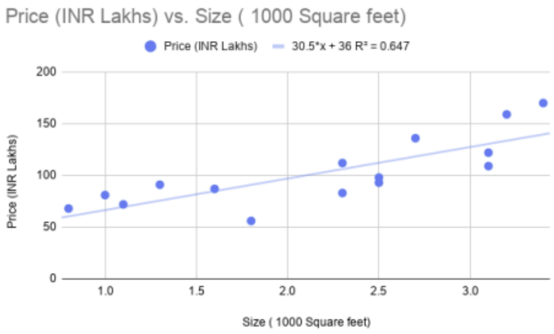
Step 1 Open the scatter plot

Step 2 Under customize tab, click on series

Step 3 Click on trendline

Step 4 Under label tab, click on use equation, and click the show R ^{2} button.

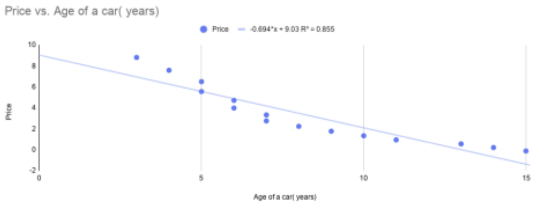
Example 1: Size versus Price of homes: Equation



Equation of the line: Price = 30.5 × Size + 36;

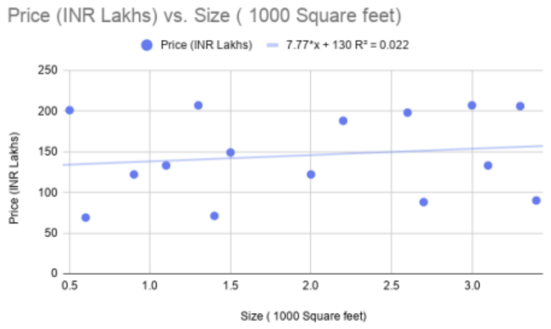
$R^2 = 0.647$; $r = 0.804$

Example 2: Age versus Price of cars: Equation



Equation of the line: Price = -0.694 × Age + 9.03; $R^2 = 0.855$; $r = -0.9247$

Example 3: Size versus Price of homes: Equation



Equation of the line: Price = 7.77 × Size + 130; R ^{2} = 0.022 ;r = 0.149

Summarizing the association with a line

I The strength of linear association between the variables was measured using the measures of Covariance and Correlation.

I The linear association can be described using the equation of a line.

Equation of line using google sheets

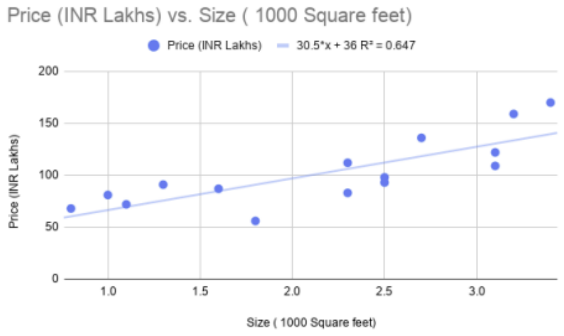
Step 1 Open the scatter plot

Step 2 Under customize tab, click on series

Step 3 Click on trendline

Step 4 Under label tab, click on use equation, and click the show R^2 button.

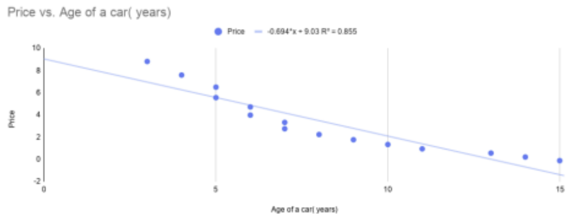
Example 1: Size versus Price of homes: Equation



Equation of the line: Price = 30.5 × Size + 36;

R ^{2} = 0.647;r = 0.804

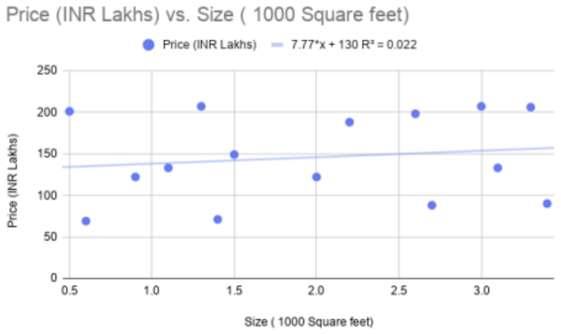
Example 2: Age versus Price of cars: Equation



Equation of the line: Price = −0.694 × Age + 9.03 ;R^2 = 0.855; r = −0.9247

Example 3: Size versus Price of homes: Equation

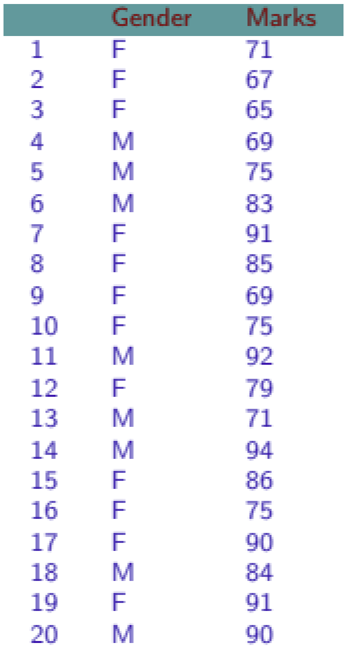
Equation of the line: Price = 7.77 × Size + 130; R^2 = 0.022; r = 0.149



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.

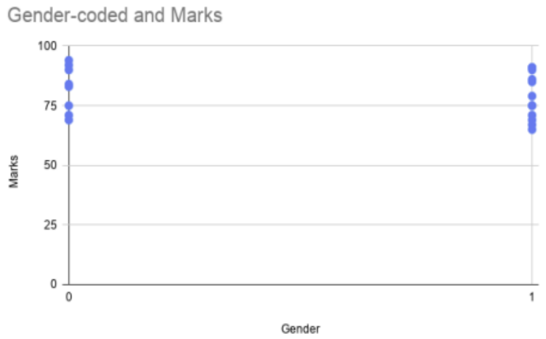
Gender versus marks-Data



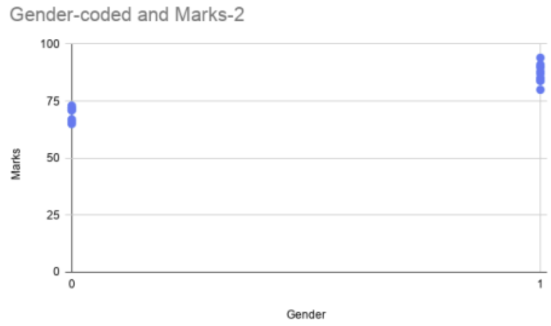
	Gender	Marks
1	F	71
2	F	67
3	F	65
4	M	69
5	M	75
6	M	83
7	F	91
8	F	85
9	F	69
10	F	75
11	M	92
12	F	79
13	M	71
14	M	94
15	F	86
16	F	75
17	F	90
18	M	84
19	F	91
20	M	90

Example 1: Scatter plot

1.



2.



Point Bi-serial Correlation Coefficient

I Let X be a numerical variable and Y be a categorical variable with two categories (a dichotomous variable).

I 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 Y_0^{-} and Y_1^{-} be the

mean values of groups with Y = 0, and Y = 1, respectively.

Step 3 Let p0 and p1 be the proportion of observations in a group with Y = 0 and Y = 1, respectively, and sX be the standard deviation of the random variable X.

The correlation coefficient

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