
# IIT MADRAS

## BSC IN PROGRAMMING & DATA SCIENCE

## NOTES W4-W6

## STATISTICS FOR DATA SCIENCE I

PROF. USHA MOHAN

NOTES BY – ADITYA DHAR DWIVEDI

## Association between two categorical variables

- Use of two-way contingency tables to understand association between two categorical variables.
- Understand association between numerical variables through scatter plots; compute and interpret correlation.
- 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: Gender versus 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 own 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 smart phone or not.
- The categorical variables in this example are
  - Gender: Male, Female (2 categories)- Nominal variable
  - Own a smartphone: Yes, No (2 categories)- Nominal variable

## Example 1: Gender versus use of smartphone-summarize data

- We have the following summary statistics
  - There are 44 female and 56 male students
  - 76 students owned a smartphone, 24 did not own.
- 34 female students owned a smartphone, 42 male students owned a smartphone.
- 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 an association.
- ✓ **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:
  - Under the Rows tab, click on the first categorical variable.
  - Under the columns tab, click on the second categorical variable.
  - 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

- A market research firm is interested in finding out whether ownership 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.
- To answer this question, a group of 100 randomly picked individuals were surveyed about whether they owned a smart phone or not.
- The categorical variables in this example are
  - Income: Low, Medium, High (3 categories) -Ordinal variable
  - Own a smartphone: Yes, No (2 categories) - Nominal variable

## Example 2: Contingency table

- We have the following summary statistics
  - There are 20 High income, 66 medium income, and 14 low-income participants.
  - 62 participants owned a smartphone, 38 did not own.
  - 18 High income participants owned a smartphone, 39 Medium income participants owned a smartphone, and 5 Low-income participants owned a smartphone.
- 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 |

**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

## Relative frequencies

### ✞ Row relative frequencies

- What proportion of total participants own a smart phone?
- What proportion of female participants own a smart phone?

| | Own a smartphone | | |
|---|---|---|---|
| Gender | 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: Row relative frequency

| | 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 |

## Example 2: Row relative frequency

| | 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 | 64.00% | 35.71% | 14 |
| Column total | 38.00% | 62.00% | 100 |

## ❖ Column relative frequencies

- What proportion of total participants are female?
- What proportion of smart phone owners are females?

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

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

## Example 1: Column relative frequency

| | 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 |

## Example 2: Column relative frequency

| | 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 |

**Section summary**

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

## 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 described earlier.
- 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.
- 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

- 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

O 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 | 64.00% | 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**

## 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 occurrence 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.

### 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

**Chart for example 1**

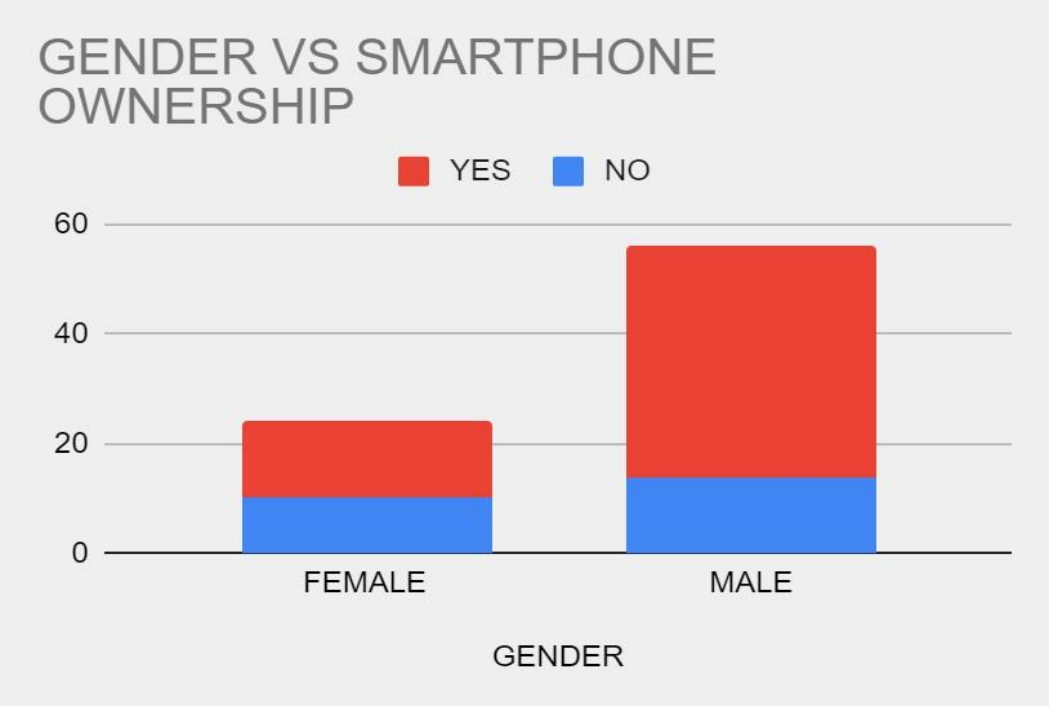
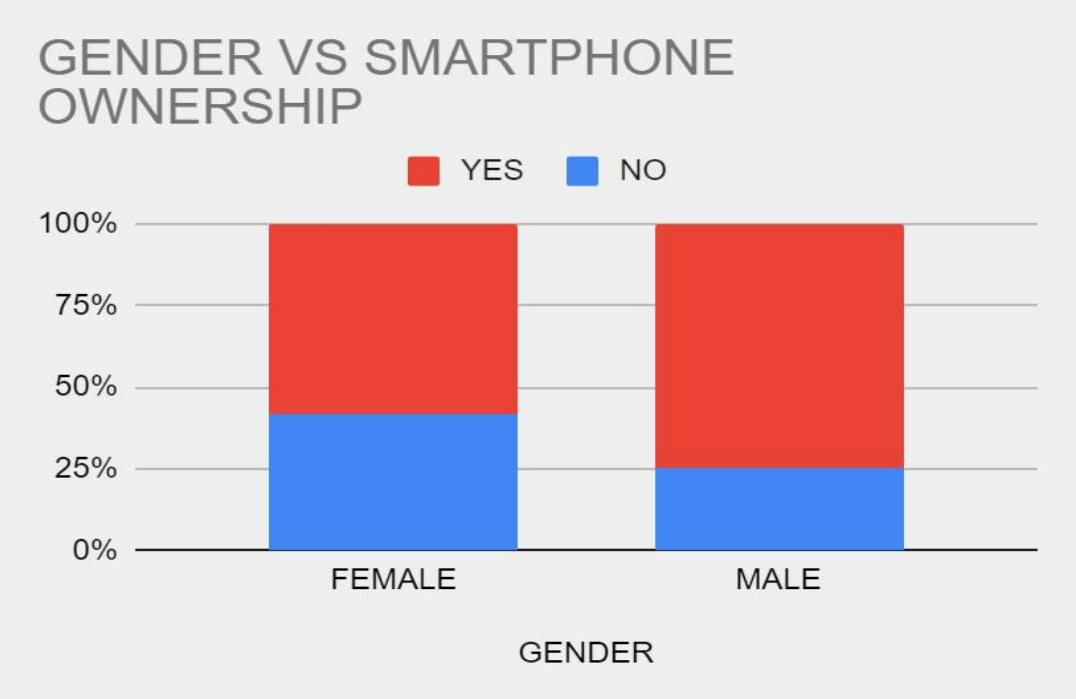



**Chart for example 2**




Section summary

- Understand whether two categorical variables are associated using the concept of relative frequencies.
- Graphical summary of association using stacked bar chart.

## Association between two numerical variables-Fitting a line

### Scatterplot

- 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

- We use a scatterplot 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.**
- 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 1: 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?

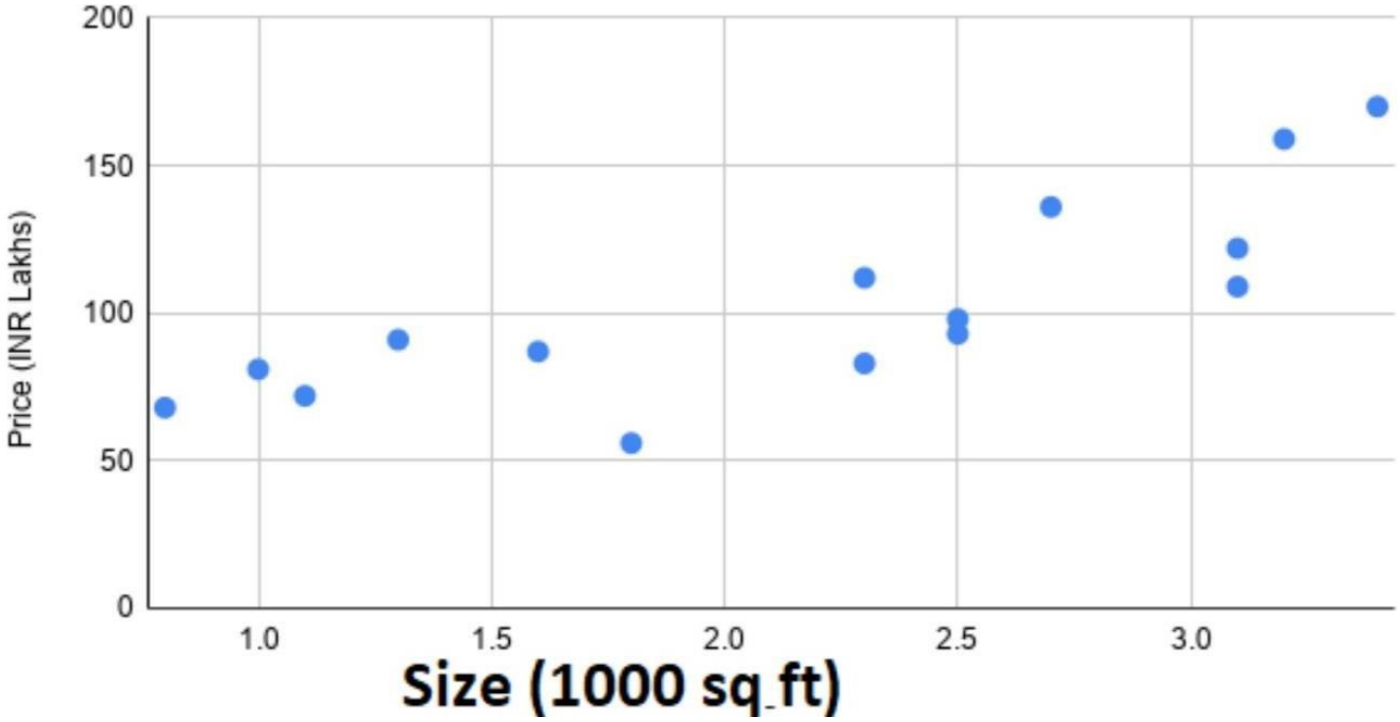
- To answer the question, he collected data on 15 homes. The data he recorded was
  - Size of a home measured in 1000 of square feet.
  - Price of a home measured in lakh of rupees.

| | 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

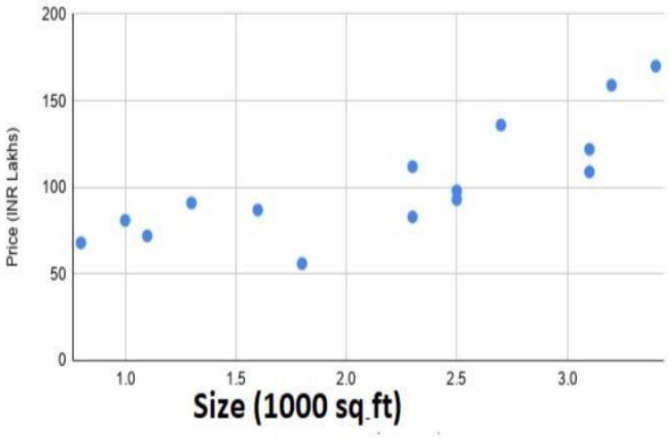
- ✓ **Step 1**: Highlight data you want to plot
- ✓ **Step 2**: Insert - chart- choose scatter chart
- ✓ **Step 3**: Under X–axis tab, choose your explanatory variable.
- ✓ **Step 4**: Under series tab, the response variable. Step 5: Label the title of the chart, axes appropriately.

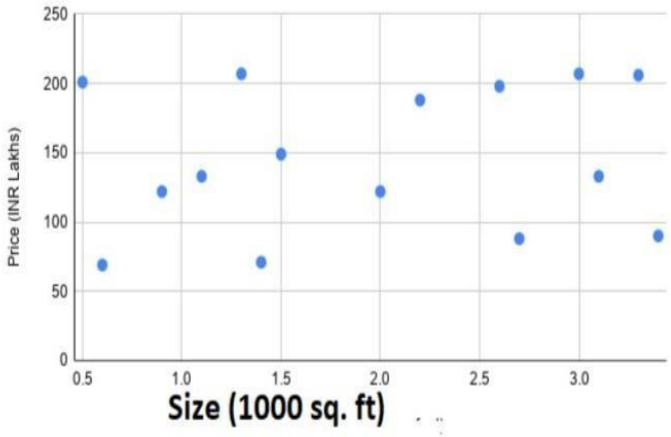
## Scatter plot




## 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?




## Section summary

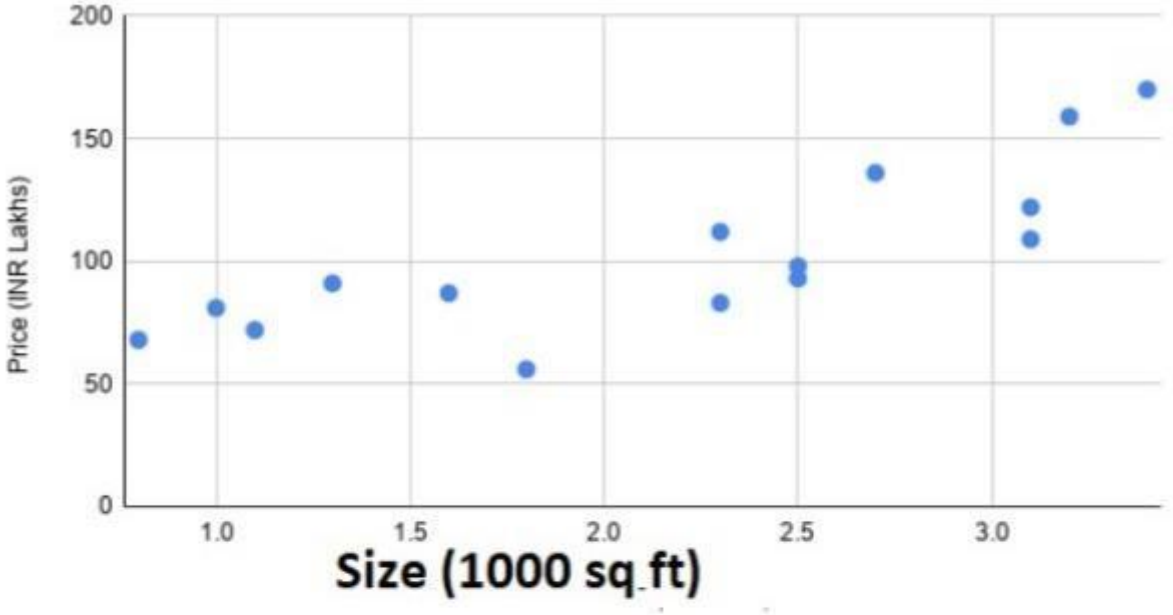
- Draw a scatter plot
- Notion of explanatory variable and response variable.
- Visual test for association

## Describing association

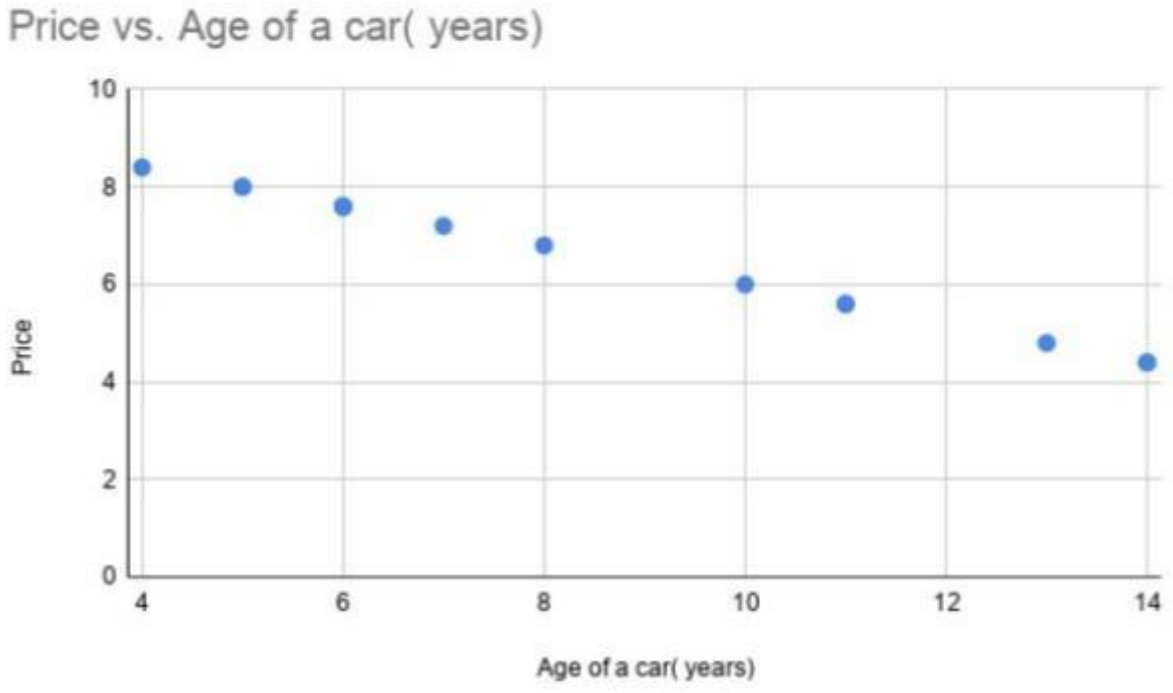
- When describing association between variables in a scatter plot, there are four key questions that need to be answered
    - Direction: Does the pattern trend up, down, or both?
    - Curvature: Does the pattern appear to be linear or does its curve?
    - Variation: Are the points tightly clustered along the pattern?
    - Outliers: Did you find something unexpected?

## Describing association: Direction

- Does the pattern trend up, down, or both?



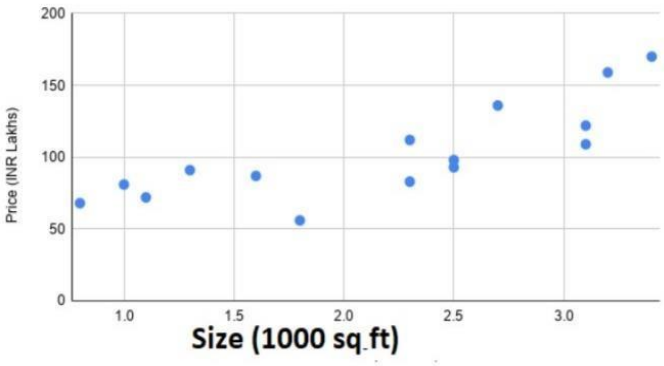

UP

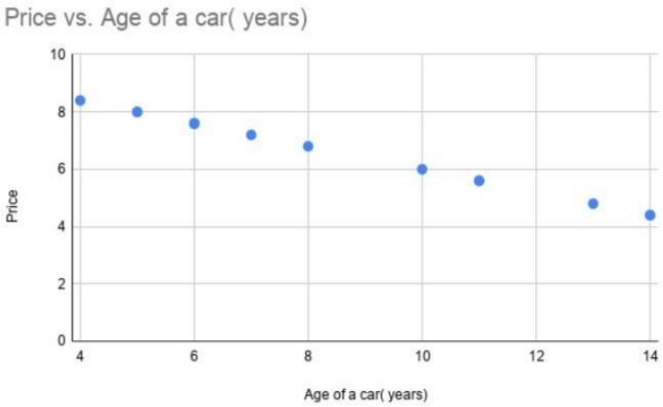


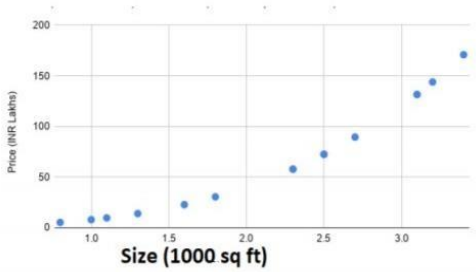

DOWN

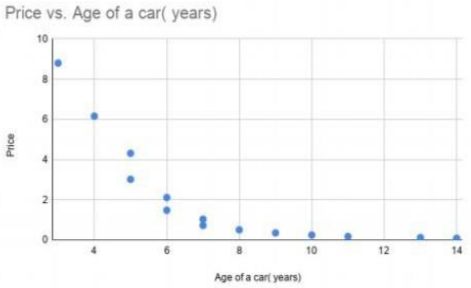
## Describing association: Curvature

- Does the pattern appear to be linear or does its curve?



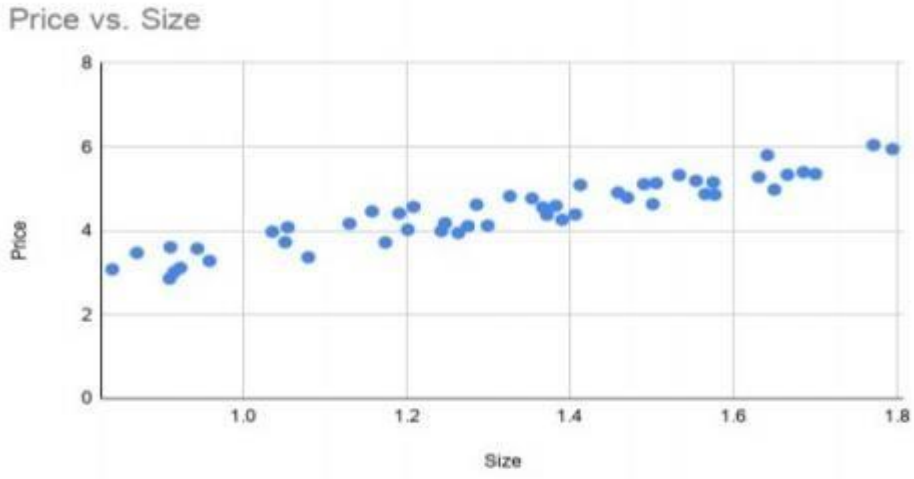



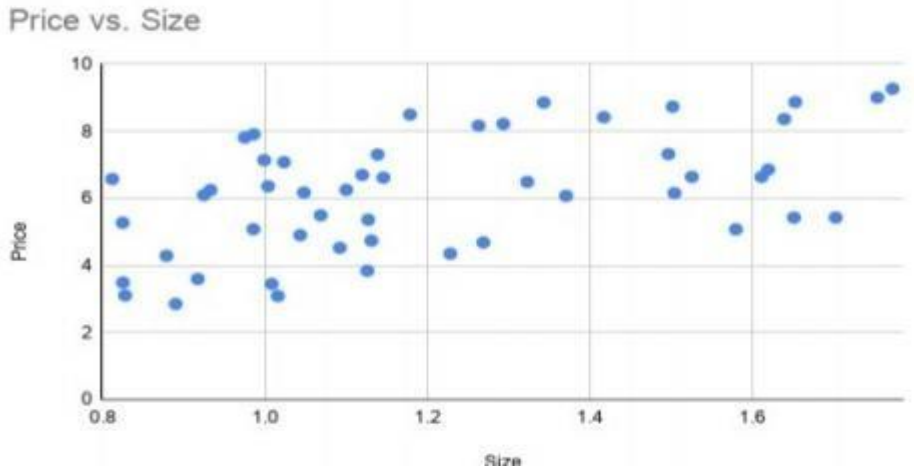




## Describing association: Variation

- Are the points tightly clustered along the pattern?



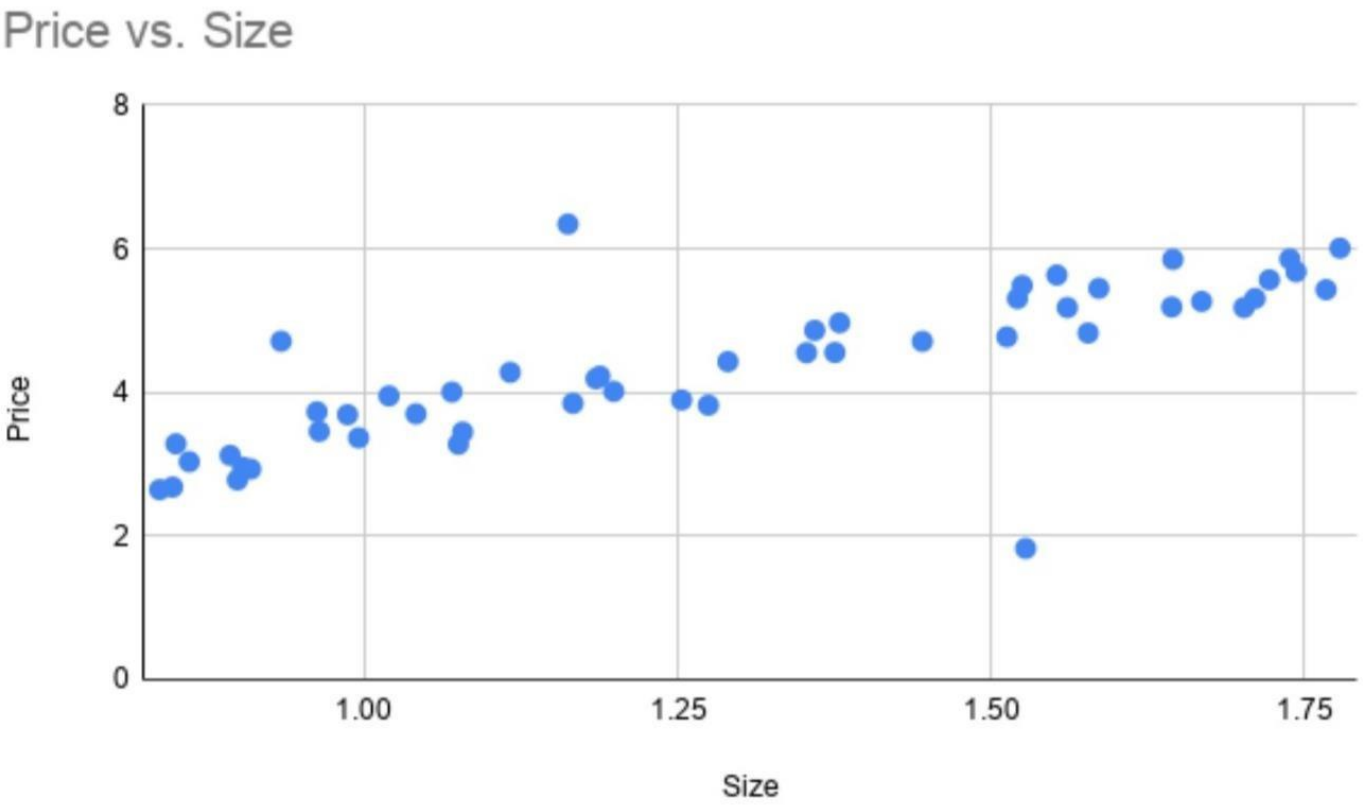

Tightly clustered




Variable

## Describing association: Outliers

- Did you find something unexpected?




## Section summary

- Describing association
  1. Direction 2. Curvature 3. Variation 4. Outliers.

## Measures of association

- How do we measure the strength of association between two variables?
  1. Covariance 2. Correlation

## ✞ Covariance

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

## Covariance: Example 1

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

| 12 | Height (CMS) y | Deviation of x $(x_i - x)$ | Deviation of y $(y_i - \bar{y})$ | $(x_i - x)\,(y_i - \bar{y})$ |
|---|---|---|---|---|
| 1 | 75 | -2 | -17.6 | 35.2 |
| 2 | 85 | -1 | -7.6 | 7.6 |
| 3 | 94 | 0 | 1.4 | 0 |
| 4 | 101 | 1 | 8.5 | 8.4 |
| 5 | 108 | 2 | 15.4 | 30.8 |

## Covariance: Example 2

- Variables: Age of a car and price of a car

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

## Key observation

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

## Covariance

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

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

## Covariance: Example 1

| Age (years) x | Height (CMS) y | Deviation of x $(x_i - x)$ | Deviation of y $(y_i - \bar{y})$ | $(x_i - x)\,(y_i - \bar{y})$ |
|---|---|---|---|---|
| 1 | 75 | -2 | -17.6 | 35.2 |
| 2 | 85 | -1 | -7.6 | 7.6 |
| 3 | 94 | 0 | 1.4 | 0 |
| 4 | 101 | 1 | 8.5 | 8.4 |
| 5 | 108 | 2 | 15.4 | 30.8 |
| | | | | 82 |

- Population covariance: 82/ 5 = 16.4
- Sample covariance: 82/ 4 = 20.5

## Covariance: Example 2

| Age x | Price y | Deviation of x $(x_i - x)$ | Deviation of y $(y_i - \bar{y})$ | $(x_i - x)\,(y_i - \bar{y})$ |
|---|---|---|---|---|
| 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: −10/ 5 = −2
- Sample covariance: −10/ 4 = −2.5

## Units of Covariance

- The size of the covariance, however, is difficult to interpret because the covariance has units.
- The units of the covariance are those of the x-variable times those of the y-variable.

**Section summary**

- Introduced the measure of covariance
- How to interpret the covariance measure

## ✞ Correlation

- A more easily interpreted 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 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

- The units of the standard deviations cancel out the units of covariance

## Remark

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

## Correlation: Example 1

| Age (years) x | Height (CMS) y | Sq. Deviation of x $(x_i - x)^2$ | Sq. Deviation of y $(y_i - \bar{y})^2$ | $(x_i - x)\,(y_i - \bar{y})$ |
|---|---|---|---|---|
| 1 | 75 | -2 | -17.6 | 35.2 |
| 2 | 85 | -1 | -7.6 | 7.6 |
| 3 | 94 | 0 | 1.4 | 0 |
| 4 | 101 | 1 | 8.5 | 8.4 |
| 5 | 108 | 2 | 15.4 | 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 x | Price y | Sq. Deviation of x $(x_i - x)$ | Sq. Deviation of y $(y_i - \bar{y})$ | $(x_i - x)\,(y_i - \bar{y})$ |
|---|---|---|---|---|
| 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 | 10 | **-10** |

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

Section summary

- Introduced measure of correlation.
- Interpreting correlation between variables.

## Summarizing the association with a line

- The strength of linear association between the variables was measured using the measures of Covariance and Correlation.
- The linear association can be described using the equation of a line.

## Equation of line using google sheets

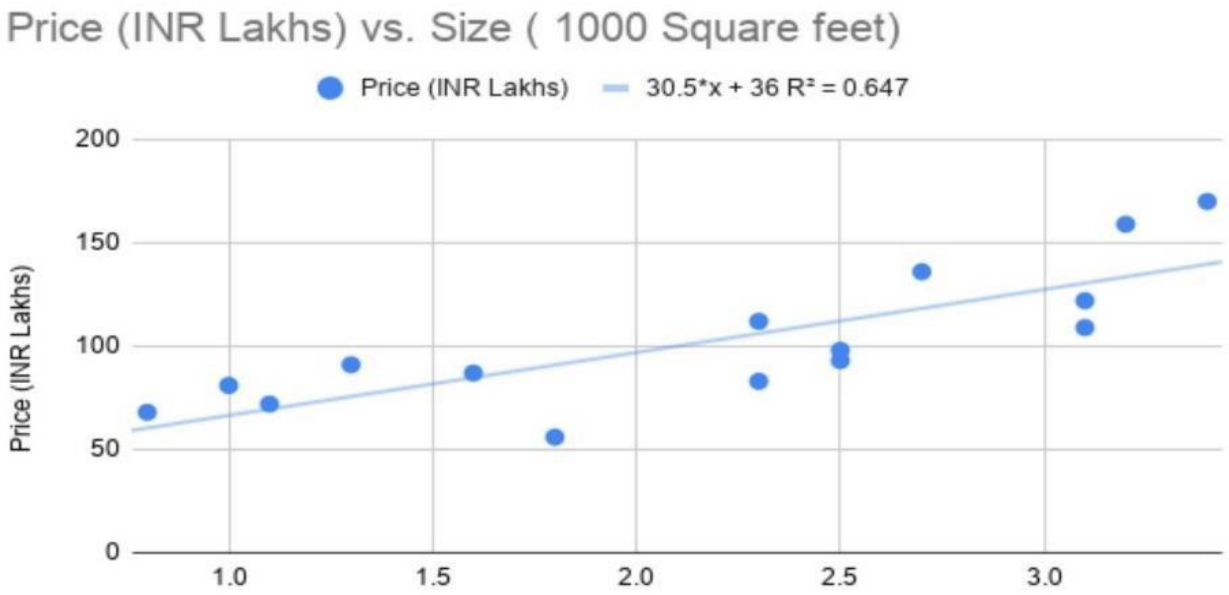
**Step 1** Open the scatter plot

**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**

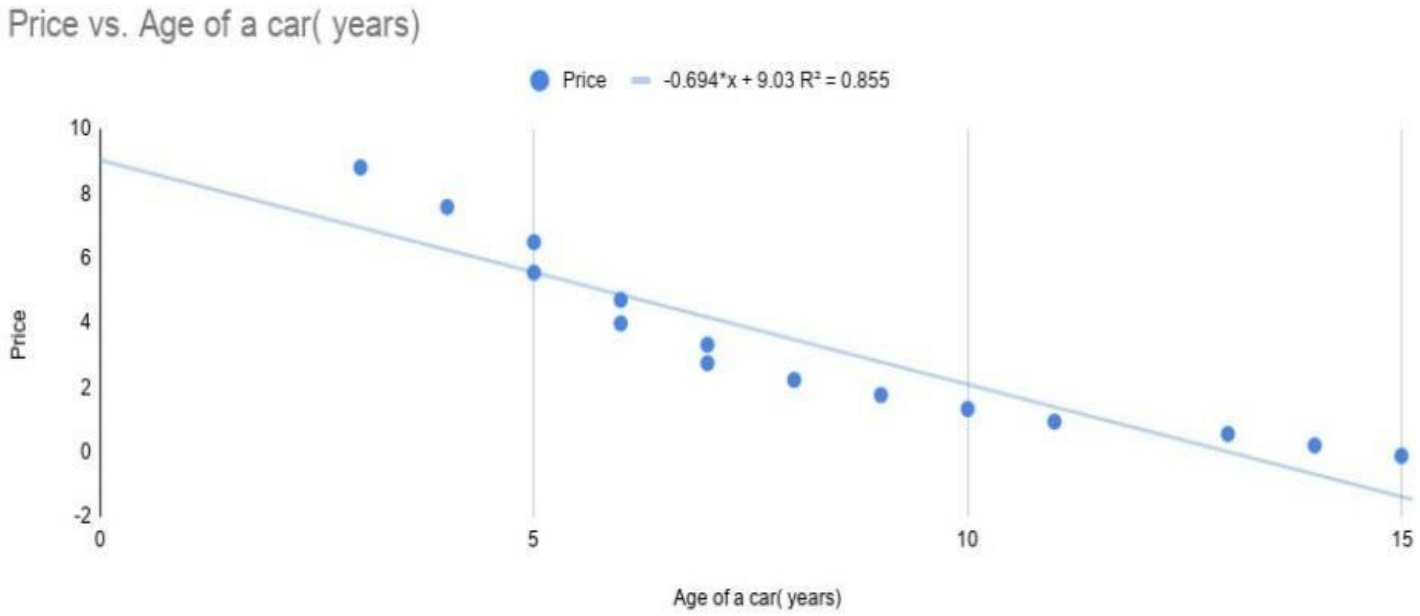
### 1: Size versus Price of homes: Equation




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

$R^2$ = 0.647;0.804

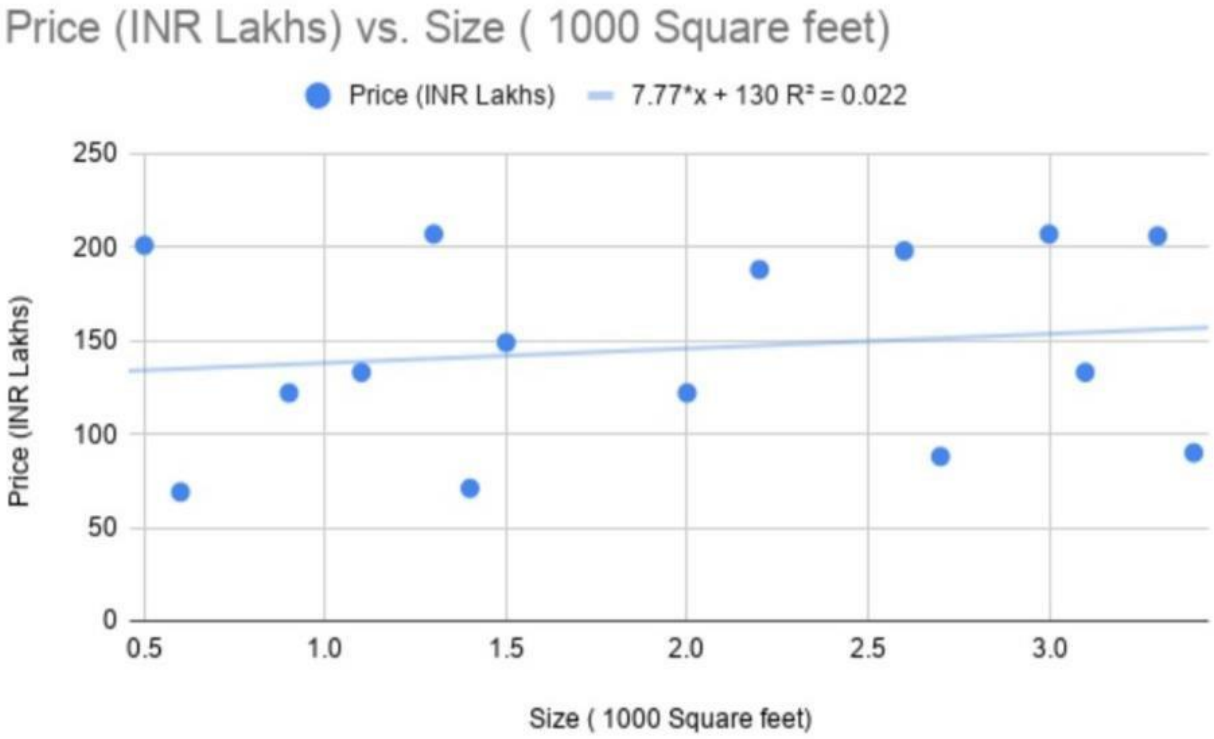
### Example 2: Age versus Price of cars: Equation




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

Section summary

- Equation of a line describing linear relationship between two variables.
- Interpreting slope, $R^2$ of the line.

### Fitting a line

- 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 and Correlation.
- 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.

**Section summary**

- Equation of a line describing linear relationship between two variables.
- Interpreting slope, $R^2$ of the line.

## Association between categorical and numerical variables

- 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 twenty students and the marks they obtained on 100 in the subject.

### Example 1: Gender versus marks-Data

| | Gender | Mark |
|---|---|---|
| 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


## 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 $\bar{Y}$ and $\bar{Y}$ 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 be the standard deviation of the random variable X.
    The correlation coefficient

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

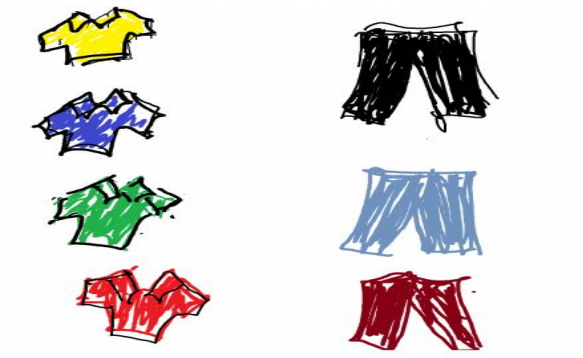
# Basic Principles of counting

## Addition rule of counting

Example 1: Buying clothes

→ You have a gift card from a major retailer which allows you to buy "one" item, either a shirt **or** a pant.

→ The choices at the retailer are___



- How many different ways can you use your card?

Solution

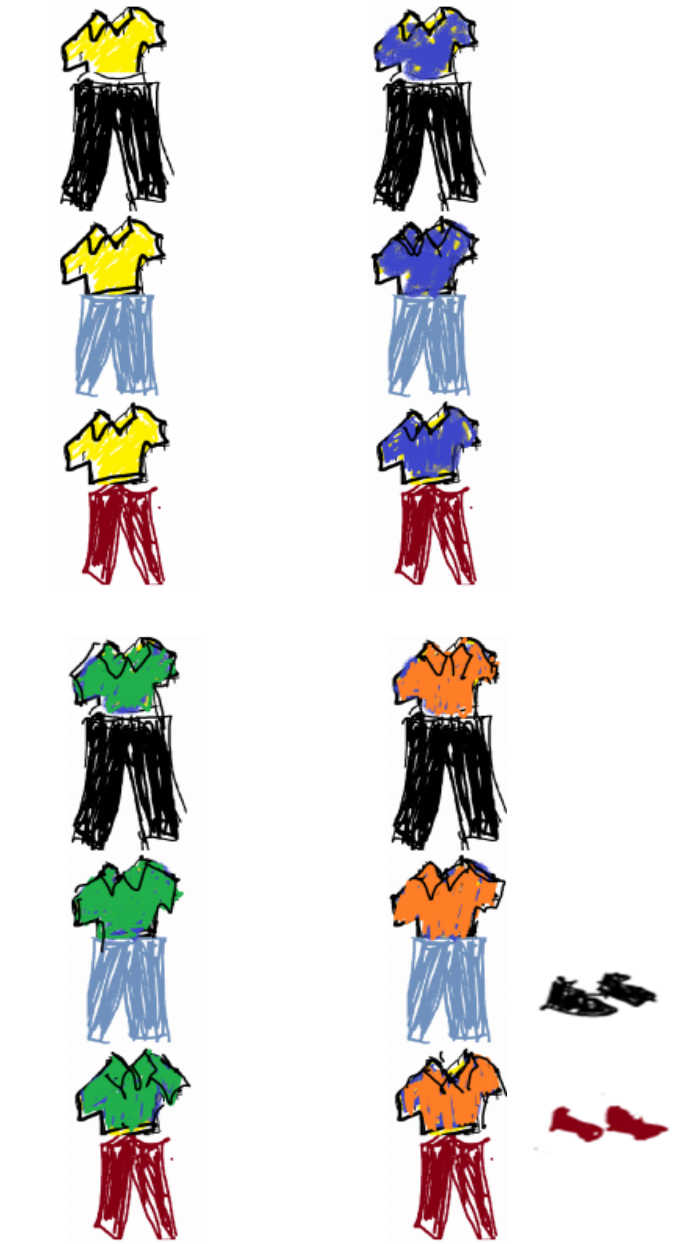
→There are four choices for buying a shirt
→There are three choices for buying a pant
→If you choose to buy a shirt (pant), you cannot buy a pant (shirt).
→Hence, the total choices available are 4 + 3 = 7

**If an action A can occur in n1 different ways, another action B can occur in n2 different ways, then the total number of occurrences of the actions A or B is n1 + n2.**

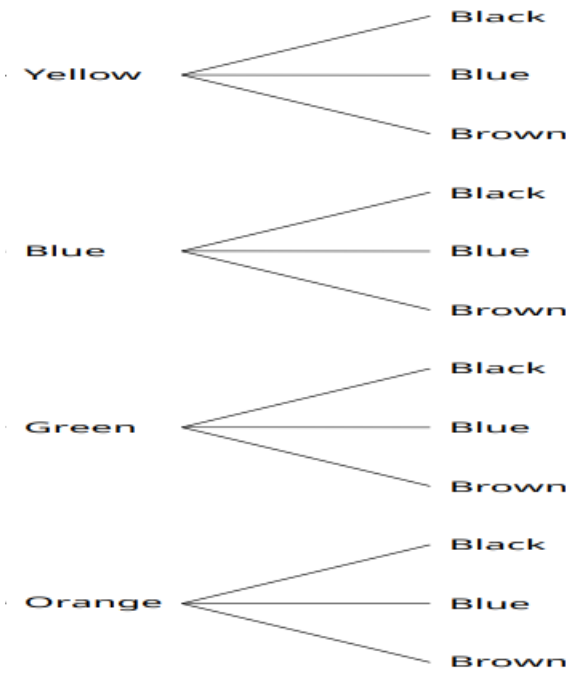
## Multiplication rule of counting

Example 2: Matching shirts and pants

→Suppose now your card allows you to buy one shirt and one pant- how many choices do you have?
→Suppose we have four shirts and three pants. How many sets can we make?



**→Tree**



Matching shirts and pants and shoes

Total 12+12= 24 ways

- **If an action A can occur in n1 different ways, another action B can occur in n2 different ways, then the total number of occurrences of the actions A and B together is $n1 \times n2$.**

- **Suppose that r actions are to be performed in a definite order. Further suppose that there are n1 possibilities for the first action and that corresponding to each of these possibilities are n2 possibilities for the second action, and so on. Then there are $n1 \times n2 \times \ldots \times nr$ possibilities altogether for the r actions.**

Example 2: Application: Creating alpha-numeric code

→Suppose you are asked to create a six-digit alpha-numeric password with the following requirement:

→The password should have first two letters followed by four numbers.

→Repetition allowed.

→Number of ways- $26 \times 26 \times 10 \times 10 \times 10 \times 10 = 6, 760, 000$

→Repetition not allowed.

→Number of ways- $26 \times 25 \times 10 \times 9 \times 8 \times 7 = 3, 276, 000$

## Factorial

Example 3: Order of finishes in a race

→There are eight athletes who take part in a 100 m race. What are the possible ways the athletes can finish the race (assuming no ties)?

→First place - any one of the 8 athletes; second - any one of the remaining 7, and so on, the seventh place – any one of the remaining 2, and finally the last place goes to the only one remaining.

→Hence the total number of ways = $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 40, 320$

- **The product of the first n positive integers (counting numbers) is called n factorial and is denoted n!**
  **In symbols,**

$$n! = n \times (n - 1) \times \ldots \times 1$$

**Remark**

→By convention $0! = 1$

Example 4:

1. $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$
2. Observe $5! = 5 \times 4!$
   → In general,
   $$n! = n \times (n - 1)!$$
3. Observe $5! = 5 \times 4! = 5 \times 4 \times 3!$
   →In general, for $i \leq n$ we have,

$$n! = n \times (n - 1). \ldots \times (n - i + 1) \times (n - i)!$$

Example 5: Simplifying expressions

1. $\frac{6!}{3!} = \frac{6\times5\times4\times3!}{3!} = 6 \times 5 \times 4 = 120$
2. $\frac{6!\times5!}{3!\times4!} = \frac{6\times5\times4\times3!}{3!}\frac{5\times4!}{4!} = 6 \times 5 \times 4 \times 5 = 600$
3. Express $25 \times 24 \times 23$ in terms of factorials-

$$\frac{25 \times 24 \times 23 \times 22 \times \ldots \times 1}{22 \times 21 \times \ldots \times 1} = \frac{25!}{22!}$$

## Permutation

- A permutation is an ordered arrangement of all or some of n objects.

### Permutation when objects are distinct

Example

Take A, B, C- Possible arrangements- taking all at a time

| First place | Second place | Third place |
|---|---|---|
| A | B | C |
| A | C | B |
| B | A | C |
| B | C | A |
| C | A | B |
| C | B | A |

Take A, B, C- Possible arrangements- taking two at a time

| First place | Second place |
|---|---|
| A | B |
| A | C |
| B | A |
| B | C |
| C | A |
| C | B |

Take A, B, C, D- Possible arrangements-

➔ Taking all at a time

| First place | Second place | Third place | Fourth place |
|---|---|---|---|
| A | B | C | D |
| A | B | D | C |
| A | C | B | D |
| A | C | D | B |
| A | D | B | C |
| A | D | C | B |
| B | A | C | D |
| B | A | D | C |
| B | C | A | D |
| B | C | D | A |
| B | D | A | C |
| B | D | C | A |
| C | A | B | D |
| C | A | D | B |
| C | B | A | D |
| C | B | D | A |
| C | D | A | B |
| C | D | B | A |
| D | A | B | C |
| D | A | C | B |
| D | B | A | C |
| D | B | C | A |
| D | C | A | B |
| D | C | B | A |

➔ Taking two at a time

| First place | Second place |
|---|---|
| A | B |
| A | C |
| A | D |
| B | A |
| B | C |
| B | D |
| C | A |
| C | B |
| C | D |
| D | A |
| D | B |
| D | C |

## Permutation formula

The number of possible permutations of r objects from a collection of n distinct objects is given by the formula

$$n \times (n-1) \times \ldots \times (n-r+1)$$

and is denoted by ${}^nP_r$

$${}^nP_r = \frac{n!}{(n-r)!}$$

- Special cases

1. ${}^nP_0 = \frac{n!}{(n-0)!} = \frac{n!}{n!}$ =1There is only one ordered arrangement of 0 objects.
2. ${}^nP_1 = \frac{n!}{(n-1)!} = n$. There are n ways of choosing one object from n objects.
3. ${}^nP_n = \frac{n!}{(n-n)!} = \frac{n!}{0!} = n!$. We can arrange n distinct objects in n! ways- multiplication principle of counting.

Examples

Take A, B, C- Possible arrangements- taking all at a time

| First place | Second place | Third place |
|---|---|---|
| A | B | C |
| A | C | B |
| B | A | C |
| B | C | A |
| C | A | B |
| C | B | A |

$$n = 3, r = 3, {}^nP_r = \frac{n!}{(n-r)!} = \frac{3!}{0!} = 6$$

Take A, B, C- Possible arrangements- taking two at a time

| First place | Second place |
|---|---|
| A | B |
| A | C |
| B | A |
| B | C |
| C | A |
| C | B |

$$n = 3, r = 2, {}^nP_r = \frac{n!}{(n-r)!} = \frac{3!}{1!} = 6$$

Take A, B, C, D- Possible arrangements- taking all at a time

| First place | Second place | Third place | Fourth place |
|---|---|---|---|
| A | B | C | D |
| A | B | D | C |
| A | C | B | D |
| A | C | D | B |
| A | D | B | C |
| A | D | C | B |
| B | A | C | D |
| B | A | D | C |
| B | C | A | D |
| B | C | D | A |
| B | D | A | C |
| B | D | C | A |
| C | A | B | D |
| C | A | D | B |
| C | B | A | D |
| C | B | D | A |
| C | D | A | B |
| C | D | B | A |
| D | A | B | C |
| D | A | C | B |
| D | B | A | C |
| D | B | C | A |
| D | C | A | B |
| D | C | B | A |

$$n = 4, r = 4, {}^nP_r = \frac{n!}{(n-r)!} = \frac{4!}{0!} = 24$$

Take A, B, C, D- Possible arrangements- taking two at a time

| First place | Second place |
|---|---|
| A | B |
| A | C |
| A | D |
| B | A |
| B | C |
| B | D |
| C | A |
| C | B |
| C | D |
| D | A |
| D | B |
| D | C |

$$n = 4, r = 2, {}^nP_r = \frac{n!}{(n-r)!} = \frac{4!}{2!} = 12$$

Example: application

➔ From a committee of 8 persons, in how many ways can we choose a chairman and a vice chairman assuming one person cannot hold more than one position? → 8 × 7 = 56

➔ Find the number of 4-digit numbers that can be formed using the digits 1, 2, 3, 4, 5 if no digit is repeated.
→ 5 × 4 × 3 × 2 × 1 = 120 ➔ How many of these will be even? → 4

➔ Six people go to the cinema. They sit in a row with ten seats. Find how many ways can this be done if

(i) they can sit anywhere: ${}^{10}P_6 = 1,51,200$

(ii) all the empty seats are next to each other: ${}^7P_6 = 5,040$

**Permutation formula**

The number of possible permutations of r objects from a collection of n distinct objects when repetition is allowed is given by the formula

$$n \times n \times \ldots \times n$$

and is denoted by $n^r$

Example

Take A, B, C- Possible arrangements- taking all at a time. n = 3, r = 3, $n^r$ = 27

| First place | Second place | Third place |
|---|---|---|
| A | A | A |
| A | A | B |
| A | A | C |
| A | B | A |
| A | B | B |
| A | B | C |
| A | C | A |
| A | C | B |
| A | C | C |
| B | A | A |
| B | A | B |
| B | A | C |
| B | B | A |
| B | B | B |
| B | B | C |
| B | C | A |
| B | C | B |
| B | C | C |
| C | A | A |
| C | A | B |
| C | A | C |
| C | B | A |
| C | B | B |
| C | B | C |
| C | C | A |
| C | C | B |
| C | C | C |

Take A, B, C- Possible arrangements- taking two at a time

| First place | Second place |
|---|---|
| A | A |
| A | B |
| A | C |
| B | A |
| B | B |
| B | C |
| C | A |
| C | B |
| C | C |

n = 3, r = 2, $n^r$= 9

## Permutation when objects are not distinct

Circular permutations

Solving of n and r using permutation formula

### Example: Rearranging letters

→ Suppose we want to rearrange the letters in the word"DATA". How many ways can it be done?

There are three distinct letters: D, A, T.

Hence the possible arrangements taking all the four letters at a time are

| First place | Second place | Third place | Fourth place |
|---|---|---|---|
| A | D | T | A |
| A | D | A | T |
| A | T | D | A |
| A | T | A | D |
| A | A | D | T |
| A | A | T | D |
| D | A | T | A |
| D | A | A | T |
| D | T | A | A |
| T | A | D | A |
| T | A | A | D |
| T | D | A | A |

→As seen in the example, we can treat the two A's in DATA as distinct. Say, A1 and A2.

→If they are treated as distinct objects, then based on the earlier formula, total number of arrangements = 4!.

→Now A1 and A2 can be arranged among themselves in 2! ways.

→A1 and A2 are essentially the same. Hence, the total number of ways the letters in "DATA" can be arranged is $\frac{4!}{2!}$= 12

**Permutation formula**

The number of permutations of n objects when p of them are of one kind and rest distinct is equal to

$$\frac{n!}{p!}$$

## Example

- Suppose we want to rearrange the letters in the word "STATISTICS". How many ways can it be done?
- Total of ten letters of which there are five distinct letters : S,T,A,I,C.
- "S" appears 3 times; "T" appears 3 times, "A' once, "I" twice, and "C" once

**Permutation formula**

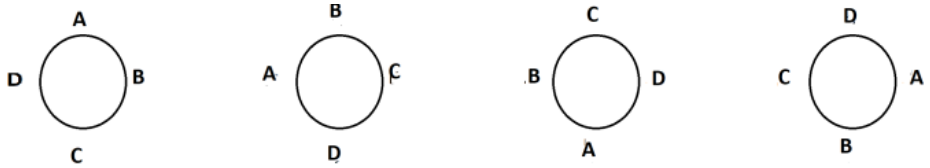
- The number of permutations of n objects where $p_1$ is of one kind, $p_2$ is of second kind, and so on $p_k$ of $k^{th}$ kind is given by $\frac{n!}{p_1!p_2!\ldots\ldots\ldots\ldots p_k!}$
- Applying the above formula to the word "STATISTICS" ;n = 10, p1 = 3, p2 = 3, p3 = 1, p4 = 2, p5 = 1.
- Hence, total number of ways = $\frac{10!}{3!3!1!2!1!}$= 50, 400

## Example

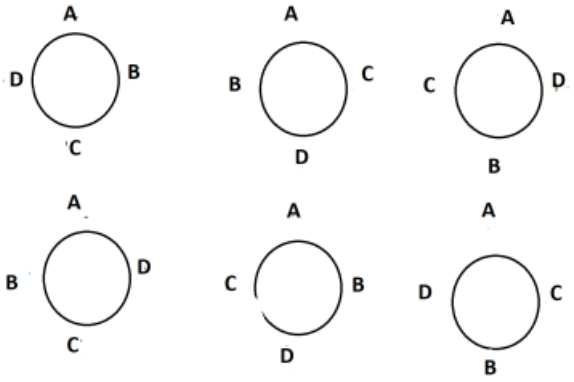
- How many ways can four people sit in a round table?
- We consider two cases: each selection is called a combination of 3 different objects taken 2 at a time.
  - Clockwise and anticlockwise are different
  - Clockwise and anticlockwise are same.

## Circular permutations Circular permutation: Clockwise and anticlockwise are different

- Consider the linear permutations of A, B, C and D
- The arrangements ABCD, BCDA, CDAB, and DABC are different when the people are seated in a row.
- However, when they are seated in a circle as shown below:

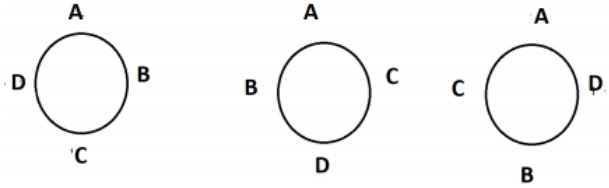



**➔Circular permutation: Clockwise and anticlockwise are different**




The number of ways n distinct objects can be arranged in a circle (clockwise and anticlockwise are different) is equal to (n − 1)!

**➔Circular permutation: Clockwise and anticlockwise are same**




The number of ways n distinct objects can be arranged in a circle (clockwise and anticlockwise are same) is equal to $\frac{(n-1)!}{2}$

**Example: Solving for n**

Find value of $n$ if ${}^nP_4 = 20\,{}^nP_2$
Answer: $\frac{n!}{(n-4)!} = 20 \times \frac{n!}{(n-2)!}$
Solving $(n-2) \times (n-3) = 20$, we get $n = -2$ or $n = 7$.
Eliminating $n = -2$, we get $n = 7$.

$\frac{{}^nP_4}{{}^{n-1}P_4} = \frac{5}{3}$
Answer: $\frac{n!}{(n-4)!} \times \frac{(n-5)!}{(n-1)!} = \frac{5}{3}$
$\frac{n}{(n-4)} = \frac{5}{3}$
Solving for $n$ gives us $n = 10$.

**Example: Solving for r**

Find $r$, if ${}^5P_r = 2.{}^6P_{r-1}$

Answer: $\frac{5!}{(5-r)!} = 2.\frac{6!}{(7-r)!}$

$\frac{5!}{(5-r)!} = 2.\frac{6!}{(7-r)(6-r)(5-r)!}$

Solving $(7-r)(6-r) = 12$ gives $r = 10$ or $r = 3$.

Since r ≤ n, the option r = 10 is eliminated and we get r = 3.

## Combinations

### Introduction

- Example: How many ways can we select two students from a group of three students?
  - Let A, B, and C be the three students.
  - We can choose AB, AC, or BC.
  - Note, when we talked of permutations, the order was important, i.e., AB was different from BA.
  - In this case, they are the same- order is not important.
- Each selection is called a combination of 3 different objects taken 2 at a time.
- In this case, the concern is only which of the 2 objects are chosen and not in the order in which they are chosen.

### Example

Consider A, B, C- Possible combinations- taking two at a time

| First place | Second place |
|---|---|
| A | B |
| A | C |
| B | C |

- Note each combination gives rise to 2! arrangements.
- All combinations give 3 × 2 = 6 arrangements.
- Number of combinations × 2! = Number of permutations

### Combinations: Notation and formula

- In general, each combination of r objects from n objects can give rise to r! arrangements.
- The number of possible combinations of r objects from a collection of n distinct objects is denoted by ${}^nC_r$ and is given by

$$ {}^nC_r = \frac{n!}{r!(n-r)!} $$

- Another common notation is $\binom{n}{r}$ which is also referred to as the binomial coefficient.

### Some useful results

1. ${}^nC_r = \frac{n!}{r!(n-r)!} = \frac{n!}{(n-r)!r!} = {}^nC_{(n-r)}$

   In other words, selecting r objects from n objects is the same as rejecting n − r objects from n objects.
2. ${}^nC_n = 1$ and ${}^nC_0 = 1$ for all values of $n$
3. ${}^nC_r = {}^{n-1}C_{r-1} + {}^{n-1}C_r; 1 \leq r \leq n$

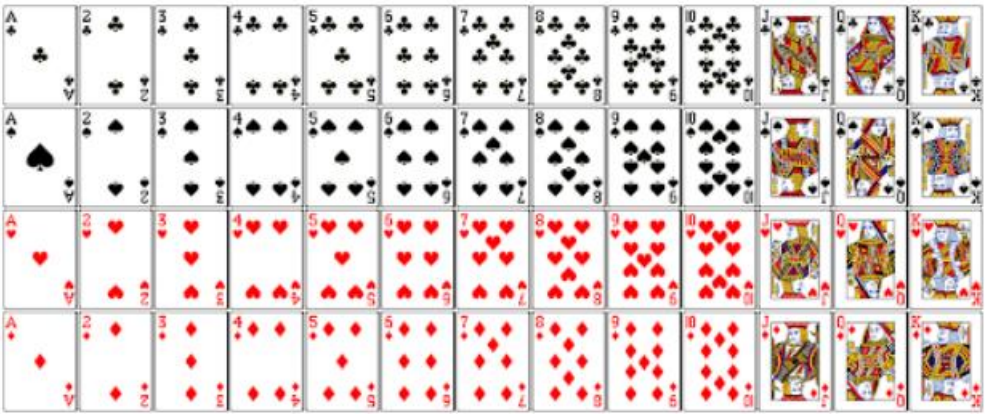
**Example**: **Choosing questions in an exam**

- In an examination, a question paper consists of 12 questions divided into two parts i.e., Part I and Part II, containing 7 and 5 questions, respectively. A student is required to attempt 8 questions in all, selecting at least 3 from each part. In how many ways can a student select the questions?

- Solution: ${}^7C_3{}^5C_5 + {}^7C_4{}^5C_4 + {}^7C_5{}^5C_3 = 35 + 175 + 210 = 420$

**Example: Game of cards**

Let's consider the case of choosing four cards from a deck of 52 cards.



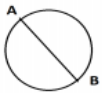
1. Total number of ways of choosing four cards from 52 cards = ${}^{52}C_4 = \frac{52!}{4!48!} = 2,70,725$
2. All four cards are of the same suit ${}^{4}C_1 \times {}^{13}C_4 = 4 \times \frac{13!}{4!9!} = 2860$
3. Cards are of same colour ${}^{2}C_1 \times {}^{26}C_4 = 2 \times \frac{26!}{4!22!} = 2,99,00$

**Example: Choosing a cricket team**

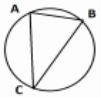
- Select a cricket team of eleven from 17 players in which only 5 players can bowl. The requirement is the cricket team of 11 must include exactly 4 bowlers? How many ways can the selection be done?
- Solution:
  - Total number of players available for selection: 17
    Number of bowlers: 5
  - Need four bowlers: This selection can be done in ${}^{5}C_4$ ways.
  - Remaining seven players can be selected from remaining twelve players in ${}^{12}C_7$ ways.
  - Total number of ways the selection can be done is ${}^{5}C_4 \times {}^{12}C_7 = 5 \times 792 = 3960$ ways.

**Example: Drawing lines in a circle**

- Given n points on a circle, how many lines can be drawn connecting these points?
- n = 2 points, one line can be drawn connecting the point

  line segment: AB
- n = 3 points, three line can be drawn connecting the points

  line segments: AB, AC, and BC
- In general, given n points, number of line segments that can be drawn connecting the points is ${}^{n}C_2$

## Applications: Permutations or combinations

- Important to distinguish between situations involving combinations and situations involving permutations.
- Permutation- "order matters". Combination - "order does not matter"

**Example: Finishing a race**

- Consider the situation of eight athletes participating in a 100m race in a competition with several rounds.
  1. How many different ways can you award the Gold, Silver, and Bronze medals?
  2. How many different ways can you choose the top three athletes to proceed to the next round in the competition? I
- Solution:
  1. How many different ways can you award the Gold, Silver, and Bronze medals?
     Order is important- Hence we need permutation. Answer is ${}^{8}P_3$ = 336 ways.

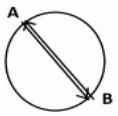
2. How many different ways can you choose the top three athletes to proceed to the next round in the competition?
   Order is not important- Hence we need combination. Answer is ${}^{8}C_{3}$ = 56 ways

**Example: Selecting a team**

- Consider the situation of a class with forty students.
  1. How many different ways can we choose two leaders?
  2. How many different ways can we choose a captain and vice-captain?
- Solution:
  1. How many different ways can we choose two leaders?
     Order not important- -hence, combination Answer: ${}^{40}C_{2}$ = 780 ways
  2. How many different ways can we choose a captain and vice-captain?
     Order important- -hence, permutation Answer: ${}^{40}P_{2}$ = 1560 ways

**Example: Drawing lines in a circle**

- Given n points on a circle, how many lines can be drawn connecting these points?
- Solution:
  1. If the segment has a direction line segment AB is different from BA. Order is important. Hence, total number of ways is ${}^{n}P_{2}$




  2. If segment has no direction. Line segment AB. Order is not important. Hence, total number of ways is ${}^{n}C_{2}$