UNIT I

VARIABLES FOR DATA ANALYTICS:

Types of Variables: Determine the nature of variables in data analysis - Differentiate between numerical and categorical. Variables - Distinguish between nominal and ordinal variables - Differentiate between interval and ratio - Distinguish between continuous and discrete

Q) What are Data Analytics? Explain the Role of Data Analytics Essentials in Business Organization.

A) Data analytics is **the foundation of the business decision-making process**.

Data analytics has become an essential business process for any company. Executives require in-depth knowledge of their business, the market, competitors and customers to make the right decisions and steer their business towards prosperity.

Data analytics is **the process of turning data into knowledge**, intelligence and business insights.
In the business context, **data analysis** is used to **understand the state of the business, the behaviour of customers, competitors and the market**; identify pain points, errors or unproductive strategies, define target customers and buyer personas, test or discard theories, etc.

The process involved in Data Analytics for Businesses.

Below is the procedure that must be fulfilled for the Data Analytics process:

- **Setting up your Goal:** You must determine the primary goal for which the Data Analytics is being used.
- **Data Collection:** Authentic Data needs to be collected for proper Data Analysis.
- **Data Cleaning:** It is essential to get a proper dataset for analysis without any errors or missing data.
- **Data Interpretation:** You should review the data using proper procedures so that relevant data is achieved to reach a valid conclusion.
- **Data Manipulation:** Manipulate the data using some Statistical tools like Excel, Google Sheets, Pivot Tables, etc. Try deeper analysis of data by finding correlations or outliers between datasets.
- **Data Visualization:** A Graphical representation of data provides better insights. Data can be visualized using Pie Charts, Bar Graphs, Histograms, etc.

Advantages of Data Analytics for Business

Multiple advantages are associated with Data Analytics for Businesses. Out of several advantages, some key benefits are discussed below:

- **High Performance:** Business Enterprises always seek to fulfill their tasks with the best processes and on-time delivery. This hasn't been possible with traditional approaches. It is also a bit tricky to communicate with the customers and fulfill their requirements. With the evolution of Data Analytics, Businesses have an access to better ways to deal with the requirements filling them with better quality and high performance.

Neelima

- **Consistency:** Data Analytics on several datasets helps gain in-depth information and insights to make strategic decisions. It also helps to maintain quality and consistency because of ETL (Extract, Transform, and Load) procedures. It gives quality output in the interactive form of patterns. Typically, Data Analytics should be performed over exact information so that Businesses gain quality output.
- **Ease in Accessing Data:** Data in structured form becomes more easily accessible. Users can access this anywhere and anytime. It improves data efficiency and even helps the group of associates/team members to access appropriate data without any hassle.

Key Types of Data Analytics for Business

To perform effective Data Analytics for Business, it is very important to understand some key terms to gauge high-level Insights. There are four key terms that Data Analysts must understand. These are as follows:

- Descriptive Analytics
- Diagnostic Analytics
- Prescriptive Analytics
- Predictive Analytics

Descriptive Analytics

Descriptive Analytics is the process in which the Business applies techniques to interpret data using various tools like Excel, SPSS (Statistical Package for the Social Sciences), Stata, MATLAB (MATrix LABoratory), etc. For instance, Businesses can spot the trending patterns based on Sales data using Descriptive Analytics.

Diagnostic Analytics

Diagnostic Analytics answers questions raised by the datasets defined in the previous step. For instance, the reason for decreasing Sales, the selling of particular products, etc. These questions that Businesses discovered from Descriptive Analytics are answered by Diagnostic Analytics.

Prescriptive Analytics

This type of Analytics preferably relies on some advanced Machine Learning Algorithms and Big Analytics to interpret data on a real-time basis. Prescriptive Analytics defines the accuracy of data based on the previous dataset. For instance, a cab booking app connects the nearest driver to customers through Google Maps. This is possible due to Prescriptive Analytics.

Predictive Analytics

Predictive Analytics leverages the previous dataset received from Diagnostics Analytics and Descriptive Analytics. This type of Analytics utilizes the previous datasets to train the models to get better outcomes. With the help of Predictive Analytics, professionals can interpret data and build predictive models.

Data Analytics Tools

- **Microsoft Excel** is a software program that enables you to organize, format, and calculate data using formulas within a spreadsheet system. Microsoft Excel may be used by data analysts to run basic queries and to create pivot tables, graphs, and

Neelima

charts. Excel also features a macro programming language called Visual Basic for Applications (VBA).

- **Python:** Python is an object-oriented open-source programming language. It supports a range of libraries for data manipulation, data visualization, and data modeling.
- **R:** R is an open-source programming language majorly used for numerical and statistical analysis. It provides a range of libraries for data analysis and visualization.
- **Tableau**: It is a simplified data visualization and analytics tool. This helps you create a variety of visualizations to present the data interactively, build reports, and dashboards to showcase insights and trends.
- **Power BI**: Power BI is a business intelligence tool that has an easy 'drag and drop functionality. It supports multiple data sources with features that visually appeal to data. Power BI supports features that help you ask questions to your data and get immediate insights.
- **QlikView**: QlikView offers interactive analytics with in-memory storage technology to analyze vast volumes of data and use data discoveries to support decision making. It provides social data discovery and interactive guided analytics. It can manipulate colossal data sets instantly with accuracy.
- **Apache Spark:** Apache Spark is an open-source data analytics engine that processes data in real-time and carries out sophisticated analytics using SQL queries and machine learning algorithms.
- **SAS:** SAS is a statistical analysis software that can help you perform analytics, visualize data, write SQL queries, perform statistical analysis, and build machine learning models to make future predictions.

Q) What is Statistics in Business? How Statistics Can Help Your Business Succeed?

A) Originally the word 'statistics' was used for the collection of data concerning states both historical and descriptive. Now it has acquired a much wider meaning and is used for all types of data and methods for the analysis of the data. Thus, in recent times it is used in two senses, namely, singular and plural.

Statistics in business is the application of statistical methods to analyze and interpret data generated by businesses. This data can include sales figures, customer feedback, website traffic, and many other types of information. By using statistical analysis, businesses can make better decisions, optimize processes, and increase profitability.
Business statistics play a vital role in tax preparation, budgeting, proposals, and data mining. Definitely, business statistics involve aspects of mathematics, management, and market research.

It involves collecting, classifying, summarizing, organizing, analyzing, and interpreting data. The main objective of Business Statistics is to make inferences about certain

Neelima

characteristics of a population in the business domain whether the population is people, objects, or collections of information.

Types of Statistics

The two main branches of statistics are:

- Descriptive Statistics
- Inferential Statistics

Descriptive Statistics – Through graphs or tables, or numerical calculations, descriptive statistics uses the data to provide descriptions of the population.

Inferential Statistics – Based on the data sample taken from the population, inferential statistics makes the predictions and inferences.

Characteristics of Statistics

The important characteristics of Statistics are as follows:

- Statistics are numerically expressed.
- It has an aggregate of facts
- Data are collected in systematic order
- It should be comparable to each other
- Data are collected for a planned purpose

Importance of Statistics

The important functions of statistics are:

- Statistics helps in gathering information about the appropriate quantitative data
- It depicts the complex data in graphical form, tabular form and in diagrammatic representation to understand it easily
- It provides the exact description and a better understanding
- It helps in designing the effective and proper planning of the statistical inquiry in any field
- It gives valid inferences with the reliability measures about the population parameters from the sample data
- It helps to understand the variability pattern through the quantitative observations

Here are some ways that statistics can help your business succeed:

1. **Identify trends and patterns:** Statistical analysis can help you identify trends and patterns in your data, such as seasonal fluctuations in sales or changes in customer behavior. This information can help you make informed decisions about marketing, product development, and resource allocation.

Neelima

2. **Make data-driven decisions:** By using statistical analysis, you can make data-driven decisions that are based on evidence rather than intuition or guesswork. This can help you avoid costly mistakes and maximize your return on investment.
3. **Optimize processes:** Statistical analysis can help you identify inefficiencies in your processes and optimize them for maximum efficiency. For example, by analyzing the time it takes to complete a task, you can identify areas for improvement and implement changes that will save time and increase productivity.
4. **Improve customer satisfaction:** Statistical analysis can help you understand your customers' needs and preferences, which can help you tailor your products and services to meet their needs. This can improve customer satisfaction and increase customer loyalty.
5. **Forecast future trends:** Statistical analysis can help you forecast future trends in your industry, such as changes in consumer behavior, emerging technologies, or shifts in the economy. This information can help you make strategic decisions that will position your business for long-term success.

Q) What is a Variable? Write about types of Variables in Detail?

Or

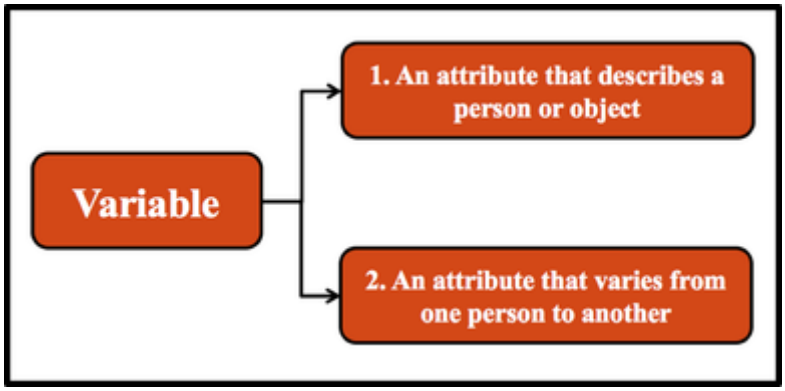
Determine the nature of Variables in data analysis

A) The definition of a variable changes depending on the context. Typically, a letter represents them, and it stands in for a numerical value. In algebra, a variable represents an unknown value that you need to find. For mathematical functions and equations, you input their values to calculate the output. In an equation, a coefficient is a fixed value by which you multiply the variable.
In statistics, a variable is a characteristic of interest that you measure, record, and analyze. Statisticians understand them by defining the type of information they record and their role in an experiment or study.

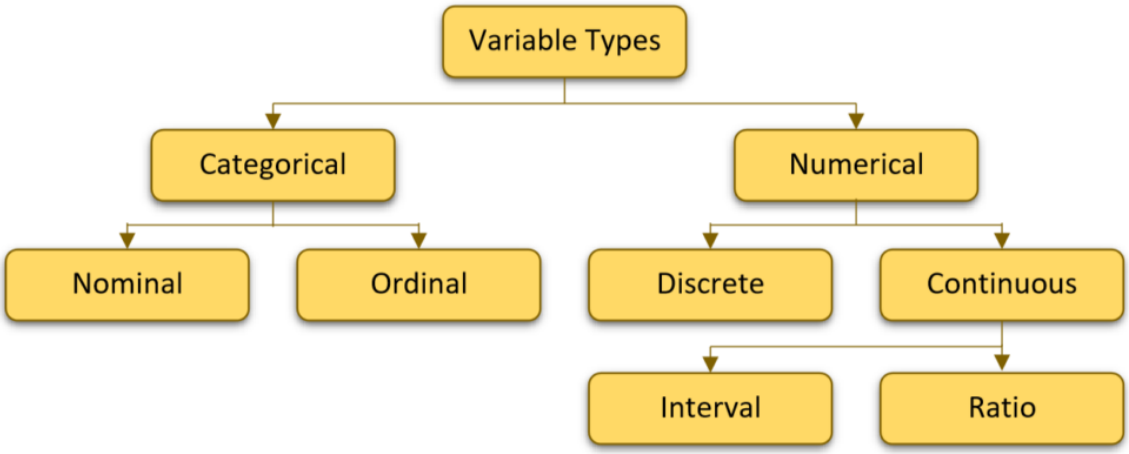
A variable is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item.

It is called a variable because the value may vary between data units in a population, and may change in value over time. For example; 'income' is a variable that can vary between data units in a population and can also vary over time for each data unit.

Neelima



Examples of variables: Age, sex, business income and expenses, country of birth, capital expenditure, class grades, and eye colour, etc.



In Statistics, there are two types of variables. They are

- **Categorical Variable or Qualitative variable**
- **Numerical Variable or Quantitative variable**

Categorical Variable or Qualitative variable

- The Categorical variable is also known as a Qualitative variable. It represents the characteristics of the data object.
- Sometimes categorical variables take numerical values, but having numerical values doesn't mean it has any mathematical meaning.
- There are two types of Categorical variables. They are
  - Nominal Variables
  - Ordinal Variables

Neelima

Data can be classified into various types in statistics based on their characteristics and properties. Ordinal data and nominal data are two examples of such data. Understanding the difference between ordinal and nominal data is important in many statistical analyses because it determines which analysis methods to use.

NOMINAL VARIABLE	ORDINAL VARIABLE
• Nominal Variables have values that are 'labels' representing some category (or class). • The values do not have any quantitative meanings. or any relative ranking or order. • Example: Gender (Male, Female, or Transgender). The values of the variable Gender do not have any order. The statistics you can derive for such variables are mode, frequency, and proportions. • Nominal Data can be summarized by mode, frequency distribution, and proportions. • Nominal data use pie charts.	• Ordinal variables have values that represent a category that has a relative order/ranking with other categories. • The values do not have any quantitative meaning but have relative ranking or order. • Example: Qualification level (Illiterate, Undergraduate, Graduate, Post Graduate). The qualification values show the relative education level between the individuals. • Ordinal data can be summarized by **mean**, median, mode, frequency distribution, and proportions. (e.g. mean of feedback rating is meaningful) • Ordinal data use a histogram or bar chart. • Sometimes ordinal variables may take numerical values. However, the difference or ratio of those values does not have any meaning. Example: Feedback rating given on a scale of 1 to 5 (1 being Below Average and 5 being Excellent). The ratio or the difference between the two values will not be meaningful.

Uses of ordinal data

This data is an important type of data in many fields and has a wide range of uses in research and practice. Here, we will discuss some key uses of this data:

- **Surveys/Questionnaires**

It is used for surveys and questionnaires due to its "ordered" nature. Statistical analysis is used to collect responses to categorize the respondents based on their responses.

- **Research**

Researchers use this type of data to collect useful information about the subject of their research. Medical researchers, for example, will need to collect data when investigating the side effects of a medication given to 50 patients.

Neelima

- **Customer service**

This data is used by businesses to improve overall customer service. When customers use a service or buy a product from a company, they are often asked to fill out a form about their experience. It will assist businesses in improving their customer service.

- **Job applications**

Employers will occasionally use a Likert scale when collecting information from job applicants during the application process. For example, when an applicant is applying for a position as a social media manager, a Likert scale may be used to determine how familiar an applicant is with Facebook, Twitter, LinkedIn, and so on.

Some more examples of Ordinal Variables

- The first, second and third person in a competition.
- Education level with values of the elementary school education, high school graduate,
  college graduate.
- When a company asks a customer to rate the sales experience on a scale of 1-10.
- When customers rank brands on the basis of their preferences.
- Pay bands in a company, as indicated by A, B, C, and D.
- Letter grades: A, B, C, and etc.
- Economic status: low, medium and high.
- When you perform a survey and ask respondents to express their level of satisfaction with the choice of those words: very satisfied, satisfied, neutral, dissatisfied, very dissatisfied.
- When a respondent should put a value from 1 to 3 to a statement. Often the words “agree, neutral, disagree” are used.

Some more Examples of Nominal Variables

- Gender (Women, Men)
- Religion (Muslin, Buddhist, Christian)
- Hair color (Blonde, Brown, Brunette, Red, etc.)
- Housing style (Ranch House, Modernist, Art Deco)
- Marital status (Married, Single, Widowed)
- Ethnicity (Hispanic, Asian)
- Eye color (Blue, Green, Brown).

Numerical Variable

- Numerical Variable is also called a Quantitative Variable.
- Numerical data is either an integer or a real value.
- It has a mathematical meaning.
- **Examples :**
  - **The census – Discrete numerical data**
  - **Measures of time – Continuous numerical data**

Neelima

- **Temperature – Continuous numerical data (grouped data)**
- **Weight – continuous numerical data**
- **Height – continuous numerical data**
- **Company revenue – discrete numerical data (ungrp/3w6`ouped)**

There are two types of Numerical Variables. They are,

- **Discrete Variable**
- **Continuous Variable**

Discrete Variable

- A Discrete Variable can only take certain distinct and separate values. But, it cannot take any intermediate value between them.

  Example: The number of transactions done by a customer on a particular day. It can 0, 1, 2, …. but it cannot be 2.5 or 2.75.

Continuous Variable

- A continuous variable can only take any continuous value.

  Example: Account Balance of the customer.

- There are two types of continuous variables. They are,
  - **Interval Variable**
  - **Ratio Variable**

INTERVAL VARIABLE	RATIO VARIABLE
1)The quantitative variables where the ratio of its values do not have any meaning and it does not have an inherently defined zero value are Interval Variables **2)**The difference between the two values of the interval variable is meaningful, however, the ratio does not make any sense. **Ex. Temperature** • Say the temperature in Kashmir is 1° C and the temperature in Mumbai is 35° C. • The ratio of Mumbai temperature by Kashmir temperature will be 35. If we make a statement that Mumbai is 35 times hotter then it would not make sense.	1) The quantitative variables which are measured on a scale such that the ratio of its values are meaningful and they have an inherently defined zero value are called Ratio Variables 2) The difference or ratio between two values for ratio variables are both meaningful **3) Ex. Length of Conference Hall** • **Width of Road A is 120 feet and the Width of Road B is 60 feet** • **The ratio of Road A width to Road B width is 2. We can say that Road A is 2 times wider than Road B and this statement is perfectly meaningful.**

Neelima

• However, if we say the difference between Kashmir and Mumbai temperature is 34° C then it would be perfectly meaningful. **3)**Interval variable does not have an inherently defined 0. • If the temperature of a particular city is 0° C then it does not mean that temperature does not exist.	• **The difference between the Road A width and Road B width is also** meaningful 4) Ratio variables have an inherently defined zero value. • If we say the width of Road C is 0. It means the road does not exist.

Pros and Cons of Interval Variable :

Advantages of Interval Scale are:

a) This scale has consistent equivalent separations between each progressive quality.
b) This scale has wide extension because of its trademark that it can sort the information in equivalent spans.
c)Mean and standard deviation can be applied to information estimated utilizing this scale. The range can likewise be determined to get the information scattering.

Disadvantages of Interval Scale are:

a) This scale has all the qualities of a total estimating scale aside from that it doesn't have an absolute zero.

Advantages of Ratio Variable

There are several advantages to using ratio variables in research:

- **Precise and Meaningful Measurements:** Ratio variables allow for precise and meaningful measurements, which are essential for accurate analysis and interpretation of data. The use of a true zero point and equal intervals between values ensures that measurements are consistent and reliable.
- **Statistical Analysis:** Ratio variables allow for a wide range of statistical analyses, including correlation, regression, and hypothesis testing. This enables researchers to identify patterns and relationships in the data, and to make informed decisions based on their findings.
- **Arithmetic Operations:** Arithmetic operations such as addition, subtraction, multiplication, and division can be performed on ratio variables, which allows for more complex calculations and comparisons.
- **Greater Discriminatory Power**: Ratio variables have greater discriminatory power than other types of variables, such as nominal and ordinal variables. This means that they can provide more detailed and accurate information about the attribute being measured.
- **Greater Range of Values:** Ratio variables have a greater range of possible values than other types of variables, which makes them useful for measuring a wide range of attributes.
- **Useful in Many Fields:** Ratio variables are used in many fields, including science, economics, finance, health research, and social sciences. This makes them a versatile tool for researchers in a variety of disciplines.

Limitations of Ratio Variable

While ratio variables have several advantages, there are also some limitations to their use in research:

- **Limited Applicability**: Ratio variables are only appropriate for measuring attributes with a true zero point and equal intervals between values. This means that they may not be useful for measuring attributes that do not meet these criteria.

Neelima

- **Cost and Time:** Obtaining precise measurements of ratio variables can be costly and time-consuming. For example, measuring physical characteristics such as weight or length may require specialized equipment and trained personnel.
- **Data Cleaning:** Ratio variables can be sensitive to outliers and missing data, which can affect the validity and reliability of the results. Data cleaning is necessary to ensure that the data is accurate and complete.
- **Inappropriate Analysis:** Inappropriate statistical analysis can lead to erroneous conclusions. For example, performing a correlation analysis on ratio variables that are not linearly related can lead to false conclusions.
- **Limited Range:** Ratio variables can have a limited range of values, which can make it difficult to measure extremely large or small values.

Example:

Nominal	Nominal	Nominal	Ordinal	Discrete	Interval	Ratio
Emp_ID	**City**	**Department**	**Designation**	**No. of Subordinates**	**Months Since Last Leave**	**Salary**
2453	Mumbai	Marketing	Vice President	4	7	125000
2589	Thane	Finance	General Manager	7	6	80000
3048	Surat	HR	Manager	10	5	50000
2985	Chennai	Operations	Asst. Manager	5	4	30000
3184	Delhi	Operations	Executive	1	1	20000
1085	Mumbai	Admin	Office Boy	0	0	8000

Q) Differentiate between Numerical and Categorical Variables.

A)

15 differences between Categorical data and Numerical data

Features	Categorical data	Numerical data

Neelima

Definition	Categorical data refers to a data type that can be stored and identified based on the names or labels given to them.	Numerical data refers to the data that is in the form of numbers, and not in any language or descriptive form.
Alias	Also known as qualitative data as it qualifies data before classifying it.	Also known as quantitative data as it represents quantitative values to perform arithmetic operations on them.
Examples	What is your gender? • Male • Female • Other	What is your test score out of 20? • Below 5 • 5-10 • 10-15 • 15-20 • 20
Types	Nominal data and Ordinal data.	Discrete data and Continuous data.
Characteristics	• No order scale • Natural language description • Can take numerical values but with qualitative properties • Can be visualized using bar charts and pie charts	• Has an ordered scale • Not use of natural language description • Takes numeric values with numeric qualities • Can be visualized using bar charts and pie charts
User-friendly design	Can include long surveys and has a chance of pushing respondents away.	Survey interaction is easy and short, hence fewer survey abandonment issues.
Data collection method	Nominal data: open-ended questions Ordinal data: multiple-choice questions	Mostly collected through multiple-choice questions and sometimes through open-ended questions.
Data collection tools	Questionnaires, surveys, and interviews	Questionnaires, surveys, interviews, focus groups and observations

Neelima

Analysis and interpretation	Median and mode Eg: univariate statistics, bivariate statistics, regression analysis	Descriptive and inferential statistics Eg: measures of central tendency, turf analysis, text analysis, conjoint analysis, trend analysis
Uses	Used when a study requires respondents' personal information, opinions and experiences. Commonly used in business research	Used for statistical calculations as a result of the potential performance of arithmetic operations
Compatibility	It is not compatible with most statistical analysis methods, hence researchers avoid using it most of the times	It is compatible with most statistical calculation methods.
Visualization	Can be visualized using only bar graphs and pie charts.	Can be visualized using bar graphs, pie charts as well as scatter plots.
Structure	Is known as unstructured or semi-structured data It can use indexing methods to structure data like Google, Bing, etc.	It is structured data and can be quickly organized and made sense of

Categorical Vs Numerical Data: Pros and Cons

Advantages

Categorical Data:

1. Provides intuitive representation of the data.
2. To gain a deeper knowledge of a topic or a population.
3. Respondent-dependent data.

Numerical Data:

1. Supports statistical calculations.
2. Gathers data quickly from a larger group.
3. Makes analysis easier.

Neelima

Disadvantages

Categorical Data:

1. Large data to process and analyze.
2. The researcher may have to handle irrelevant data.
3. Statistical analysis cannot be performed.
4. The data is likely to be low sensitive, such as yes/no or good/bad.

Numerical Data:

1. Standardized performance limits investigation.
2. Significant factors can be eliminated which alters the results.
3. The researcher has more control over the data than the respondents.
4. Can lead to doubts about the validity of the result.

Q) Difference between Discrete and Continuous Data

A)

BASIS FOR COMPARISON	DISCRETE DATA	CONTINUOUS DATA
Meaning	Discrete data is one that has clear spaces between values.	Continuous data is one that falls on a continuous sequence.
Nature	Countable	Measurable
Values	It can take only distinct or separate values.	It can take any value in some interval.
Graphical Representation	Bar Graph	Histogram
Tabulation is known as	Ungrouped frequency distribution.	Grouped frequency distribution.
Classification	Mutually Inclusive	Mutually Exclusive
Function graph	Shows isolated points	Shows connected points
Example	Days of the week	Market price of a product

The advantages and disadvantages of discrete and continuous data are presented in the table below.

Advantages	Disadvantages

Neelima

Discrete	
It is simple to read and spot patterns in discrete data.	It exclusively uses whole numbers, making it more challenging to analyze because values cannot be broken down into smaller parts or components.
Discrete data can be easily included in a wide range of graph types.	Discrete data cannot yield additional insights since it is less detailed than continuous data.
You do not need large amounts of discrete data to make inferences or valid analyses.	Data that is discrete might not be as accurate as data that is continuous.
Continuous	
Continuous data is accurate.	Working with continuous data can be confusing.
Presenting more complex data with continuous data is more practical.	Specific measurement techniques are constrained.
Few data points can be used to draw conclusions, and small samples can still yield reliable analysis.	Continuous data collection can be more expensive because it frequently requires more time.

Neelima

Central Tendency of Data: Identify the components of central tendency - Calculate mean/median/mode - Identify the steps in calculating weighted /geometric /harmonic means - Measurement and Variability: Determine core aspects of measurement and variability - Calculate range - Calculate quartiles - Calculate interquartile range - Calculate variance - Calculate standard deviation - Analyze permutation with repetition - Analyze combinations without repetition

Q) What is Data? Explain about grouped data and ungrouped data.

A) Data can be defined as a systematic record of a particular quantity. It is the different values of that quantity represented together in a set. It is a collection of facts and figures to be used for a specific purpose such as a survey or analysis. When arranged in an organized form, can be called information.

The source of data (primary data, secondary data) is also an important factor.

Grouped data and Ungrouped data

- Ungrouped data is a set of observations or measurements that are not arranged into any specific categories or groups. It is simply a list of individual values.
- Grouped data, on the other hand, is a set of observations or measurements that have been organized into groups or classes based on certain criteria. These groups or classes are known as class intervals and are used to make the data more manageable and easier to analyze.

Ungrouped Data	Grouped Data
Original data is not arranged in any particular order	Original data is arranged into groups or classes
No class intervals or frequency distributions	Has class intervals or frequency distributions
No specific method of presentation	Specific method of presentation, such as a histogram or frequency table
Difficult to identify patterns or trends	Easier to identify patterns or trends
Not suitable for statistical analysis	Suitable for statistical analysis

Ungrouped Data

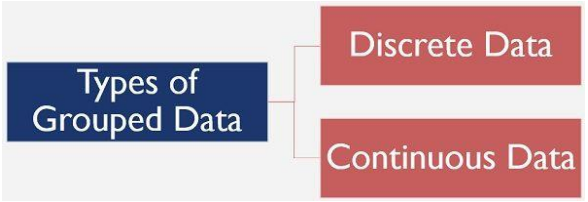
A) Ungrouped data is a type of data that has not been organized or grouped into specific categories. It is also known as raw data or unorganized data. It is simply a collection of observations or measurements without any structure or organization. It is often used as a starting point for statistical analysis, where it is grouped and organized into meaningful categories in order to draw conclusions and make inferences.

Histograms and frequency tables can be used to show this type of data

EX: Let us say there are 30 women in a colony, whose age group are as follows:
55, 35, 29, 35, 24, 77, 65, 45, 26, 29, 35, 66, 57, 59, 33, 31, 64, 28, 63, 55, 25, 69, 46, 38, 48, 61, 37, 55, 24, 64
Data available in such a form is called **raw data**. And each entry i.e., 55, 35, 29,65, and so forth, is the **value** or **observation**.

Grouped Data

Grouped data is a type of data that has been organized into specific categories or groups. It is also known as organized data or tabulated data. It is a way to present raw data in a more meaningful and manageable way. Grouped data is typically used for statistical analysis, as it allows for easier comparison and interpretation of the data.



Grouped data is of two kinds:

1. **Discrete Data**: Use of discontinuous class intervals, is called discrete data, such as 1-10, 11 – 20, etc.
2. **Continuous Data**: It is characterized by the use of continuous class limits. such as 0 – 10, 10 – 20, etc.

Frequency Distribution table

The objective of statistical interpretation is to organize data into a concise form so that interpretation and analysis become easy. It is for this reason that we organize larger data into a table called **the frequency distribution table**.

Ungrouped Data

Let the test scores of all 20 students be as follows:
23, 26, 11, 18, 09, 21, 23, 30, 22, 11, 21, 20, 11, 13, 23, 11, 29, 25, 26, 26

Marks obtained in the test	No. of students (Frequency)
09	1
11	4
13	1
18	1
20	1
21	2

22	1
23	3
25	1
26	3
29	1
30	1
Total	**20**

In the example above, the frequency refers to the number of students getting a particular mark in the test. Also, note that your frequency must always total the number of observations after tallying. Here, the total we have obtained after tallying the test scores of the students is 20 which is also the number of observations given.

A frequency distribution such as the one above is called an **ungrouped frequency distribution table**. It takes into account ungrouped data and calculates the frequency for each observation singularly.

Grouped Data

Now consider the situation where we want to collect data on the test scores of five such classes i.e. of 100 students. It becomes difficult to tally for each and every score of all 100 students. Besides, the table we will obtain will be very large in length and not understandable at once. In this case, we use what is called a **grouped frequency distribution table**.

Such tables take into consideration groups of data in the form of **class intervals** to tally the frequency for the data that belongs to that particular class interval.

Marks obtained in the test (Class Interval)	**No. of students (Frequency)**
0-5	3
5-10	11
10-15	38
15-20	34
20-25	9
25-30	5
Total	100

The first column here represents the marks obtained in class interval form. The lowest number in a class interval is called the **lower limit** and the highest number is called the **upper limit**. This example is a case of continuous class intervals as the upper limit of one class is the lower limit of the following class.

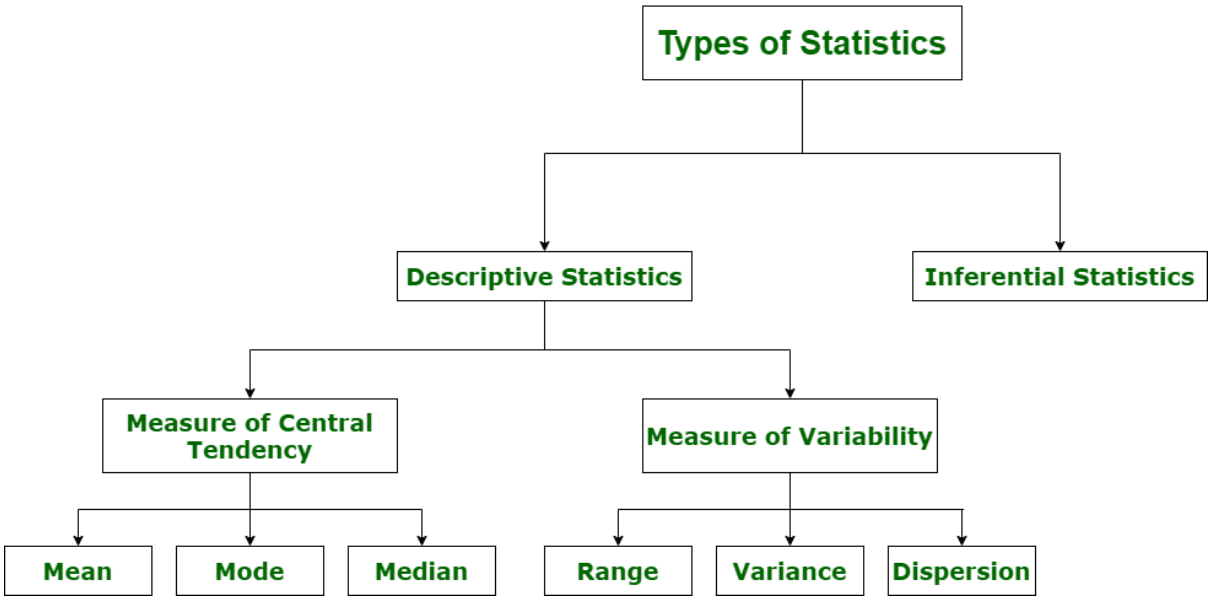
Q) What is statistics? explain types of statistics?

A) Statistics is a branch of math focused on collecting, organizing, and understanding numerical data

Statistics finds applications in diverse fields like weather forecasting, stock market analysis, insurance, betting, and data science.

There are 2 types of statistics:

- ***Descriptive Statistics***
- ***Inferential Statistics***



Descriptive Statistics

Descriptive statistics uses data that provides a description of the population either through numerical calculated graphs or tables. It provides a graphical summary of data.
It is simply used for summarizing objects, etc.

There are two categories in this as follows.

- Measure of Central Tendency
- Measure of Variability

Measure of Central Tendency

Measure of central tendency **is also known as summary statistics** that are used to represent the center point or a particular value of a data set or sample set. In statistics, there are three common measures of central tendency that are:

- Mean
- Median
- Mode

Measure of Variability

The **measure of Variability is also known as the** measure of dispersion and is used to describe **variability in a sample or population.**
In statistics, there are three common measures of variability as shown below:

- **Range**
- **Variance**
- **Dispersion**

Inferential Statistics

Inferential Statistics **makes inferences and predictions about the population based on a sample of data taken from the population.** It generalizes a large dataset and applies probabilities to draw a conclusion.

Hypothesis Testing

Types of Inferential Statistics

Various types of inferential statistics are used widely nowadays and are very easy to interpret. These are given below:

- *One sample test of difference/One sample hypothesis test*
- *Confidence Interval*
- *Contingency Tables and Chi-Square Statistic*
- *T-test or Anova*
- *Pearson Correlation*
- *Bivariate Regression*
- *Multi-variate Regression*

Q) Explain Measures of Central Tendency in understanding the performance of Business organization.

A)

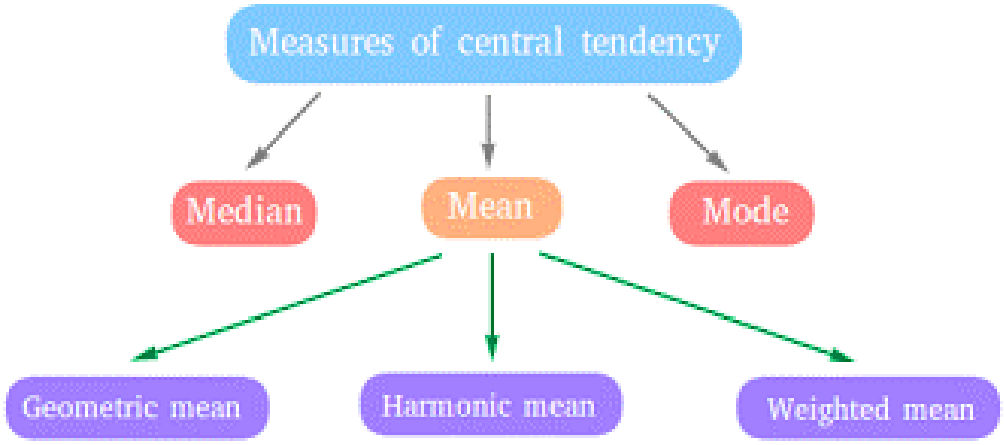
Measure of Central tendency

The representative value of a data set, generally the central value or the most occurring value that gives a general idea of the whole data set is called Measure of Central Tendency.

Components of Central Tendency

There are 3 main measures OR Components of Central Tendency are

- Mean,
- Median and
- Mode.



Mean

It is also recognized as the expected value. When we are faced with a large set of data, it helps to calculate averages to help make sense of this data. One such average is called the Mean, sometimes also known as the Arithmetic Mean. Mean is the measure of central tendency most commonly used. Mean is equal to the sum of all the values of a collection of data divided by the number of values in the data.

There are majorly 3 types of mean values that you will find in statistics.

- **Arithmetic Mean**
- **Geometric Mean**
- **Harmonic Mean**

Arithmetic Mean:

1) Arithmetic Mean calculation for ungrouped data

Mean = (Sum of all the observations/Total number of observations)

Let the data set be $x_1, x_2, x_3 \ldots\ldots x_n$

Mean = $(x_1 + x_2 \ldots\ldots x_n) / n$ or $\frac{\sum x_i}{n}$

Where n= total no. of observations.

Example :

Marks obtained by 5 students are 15,18,12,19 and 11

The Arithmetic Mean of marks obtained by students is,

Mean = $(x_1 + x_2 \ldots\ldots x_n) / n$

= (15+18+12+19+11) / 5 = 15

Q) Calculation of Descriptive statistics Mean , median , mode Using MS-Excel Functions:

How to calculate Mean in Excel?

In excel you can use the following function to find the mean or Average.

= AVERAGE(value1, [value2, ...])

The Value2 is optional. It will calculate the Mean of all the selected values in the dataset.

STEPS FOR CALCULATION OF MEAN IN EXCEL
(REFER LAB RECORD FOR PROCEDURE)

Mean

Where you can use Mean in Real Life?

- Sports – Player Average Score.
- Academics – Average Marks of the student.
- Stock Markets – Moving Average, Exponential, Simple and Weighted Moving Average.

ADVANTAGES OF MEAN

- It is easy to compute and has a unique value.
- It is based on all the observations.
- It is well defined.
- It is least affected by sampling fluctuations.
- It can be used for further statistical analysis

Disadvantages if Mean

- The mean is unduly affected by the extreme items (outliers).
- It cannot be determined for the qualitative data such as beauty, honesty etc.
- It cannot be located by observations on the graphic method.

2. Arithmetic Mean-Ungrouped Frequency Distribution

2) Arithmetic mean/ Average calculation for grouped data

When a data set is large, a frequency distribution table is often used to display the data in an organised way. A frequency distribution table lists the data values, as well as the number of times each value appears in the data set.
In a discrete frequency distribution the arithmetic mean may be computed by any one of the following methods:

- **Direct method**

Arithmetic mean by Direct method

If x_1, x_2, x_3 … are observations having frequencies f_1, f_2, f_3 , … …

$$\overline{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

Median

Median: Median of the data is the data of the middlemost observation that is achieved after modifying the data in ascending order. The advantage of applying the median as a central tendency is that it is less influenced by outliers and skewed data.

Median is determined for ungrouped data as well as grouped data.

Formula for the ungrouped data is divided into 2 categories:

$$\text{If 'n' is odd:} \quad \text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$$

$$\text{If 'n' is even:} \quad \text{Median} = \frac{\left(\frac{n}{2}\right)^{th}\text{term} + \left(\frac{n}{2}+1\right)^{th}\text{term}}{2}$$

Calculation of Median in Excel

The MEDIAN Function[1] is categorized under Excel Statistical functions. The function will calculate the middle value of a given set of numbers. Median can be defined as the middle number of a group of numbers. That is, half the numbers return values that are greater than the median, and half the numbers return values that are less than the median. For example, the median of 2, 3, 3, 5, 7, and 10 is 4.

In financial analysis, the function can be useful in calculating the median of certain numbers, e.g., median sales or median expenses.

Formula

=MEDIAN(number1, [number2], ...)

The function uses the following arguments:

1. **Number1** (required argument) – The number arguments are a set of one or more numeric values (or arrays of numeric values), for which we wish to calculate the median.
2. **Number2** (optional argument)

STEPS FOR CALCULATION OF MEDIAN IN EXCEL
(REFER LAB RECORD FOR PROCEDURE)

MERITS

- It is easy to compute. It can be calculated by mere inspection and by the graphical method
- It is not affected by extreme values.
- It can be easily located even if the class intervals in the series are unequal

DEMERITS

- It is not amenable to further algebraic treatment
- It is a positional average and is based on the middle item.
- It does not take into account the actual values of the items in the series

Median in real life

- Economics: Household Income of a Country.
- Academics: Student Placement Packages
- Real Estate: Price of a House
- Stock Market – Median Price to Earning Ratio, P/E.

For a set of **grouped data**, we can follow the following steps to find the median:

$$Median = l + \frac{\frac{n}{2} - f_0}{f_1} x\ h$$

Whereas

l = Lower limit of edian class
n = Total frequency
fo = Cumulative Frequency of class preceding the median class
f1= Frequency of median class
h= Width of median class

Mode

Mode: Mode is defined as the value which occurs most frequently in a data set. The mode obtained may be two or more in frequency distribution.

The MODE Function[1] is categorized under Excel Statistical functions. MODE will calculate the most frequently occurring value from a list of supplied numbers. If there are two or more most frequently occurring values in the supplied data, the function returns the lowest of the values.

In financial analysis, the function can be useful in calculating mode for a given dataset. For example, for a dataset of a million people, we can assign numbers to individual purchases such as TV=1, iPad=2, Cell Phone=3. The data can serve as an important and valuable marketing tool.

Using the function, we can see how frequently consumers bought each of the above gadgets.

Formula

=MODE(number1, [number2], ...)

The function uses the following arguments:

1. **Number1** (required argument) – The number arguments are a set of one or more numeric values (or arrays of numeric values) for which we want to calculate the mode.
2. **Number2** (optional argument)

For the MODE function, remember that:

- Arguments can either be numbers or names, arrays, or references that contain numbers.
- If an array or reference argument contains text, logical values, or empty cells, the values are ignored by this function. However, any cells with the value zero are included.
- Arguments that are error values or text that cannot be translated into numbers cause errors.

STEPS FOR CALCULATION OF MODE IN EXCEL
(REFER LAB RECORD FOR PROCEDURE)

Where you can use Mode in Real Life?

E-commerce- Mostly sale items

Stock Market – Price Level,(Finding the most occurrence of a Price of a stock market.)

B) For Grouped Data

General formula for mode calculation of grouped data is:

FORMULA

The mode or modal value of the distribution is that value of the variate for which the frequency is maximum. It is the value around which the items or observations tend to be most heavily concentrated. The mode is computed by the formula.

$$\text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

Modal class is the class which has maximum frequency.

f_1 = frequency of the modal class
f_0 = frequency of the class preceding the modal class
f_2 = frequency of the class succeeding the modal class
c = width of the class limits

Advantages :

- It is comparatively easy to understand.
- It can be found graphically.
- It is easy to locate in some cases by inspection.
- It is not affected by extreme values.
- It is the simplest descriptive measure of average

Disadvantages

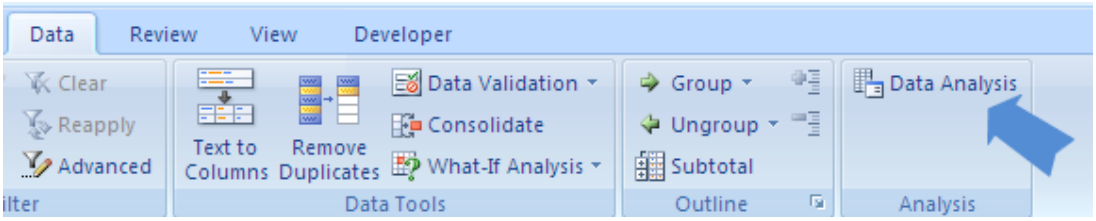
- It is not suitable for further mathematical treatment.
- It is an unstable measure as it 1S affected more by sampling fluctuations.
- Mode for the series with unequal class intervals cannot be calculated.
- In a bimodal distribution, there are two modal classes and it is difficult to determine the values of the mode

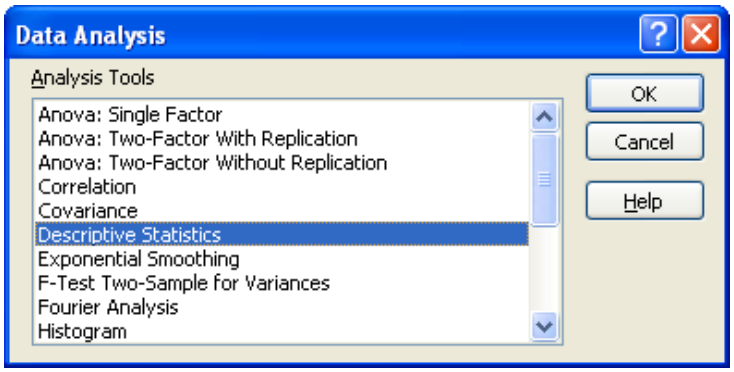
Q) How to use Descriptive Statistics in Excel?
Or
How descriptive statistics useful for data analysis

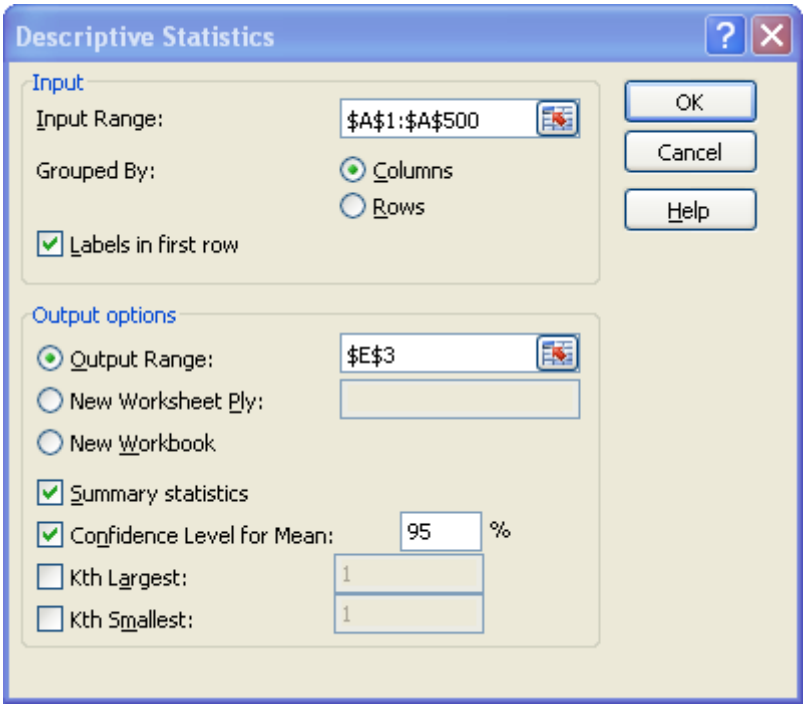
Using Excel:

1. If you haven't already installed the **Analysis ToolPak** , Click the **Microsoft Office** button, then click on the **Excel Options** , and then select **Add-Ins** , Click **Go**, check the **Analysis ToolPak** box, and click **Ok**

2. Select **Data** tab, then click on the **Data Analysis** option, then selects **Descriptive Statistics** from the list and Click **Ok**. ***[Data tab >> Data Analysis >> Descriptive Statistics]***







3. In the **Input Range** we select the data, and then select **Output Range** where you want the output to be stored. *If you don't specify the output range it will throw output in the new worksheet.*

4. Check **Summary Statistics** and **Confidence Level for Mean** options. By default the confidence level is 95%. You can change the level as per the hypothesis standard of study.

5. When you click **Ok**, you will see the result in the selected output range.

Column1	
Mean	5.533066
Standard Error	0.131332
Median	6
Mode	8
Standard Deviation	2.933741
Sample Variance	8.606836
Kurtosis	-1.27785
Skewness	-0.03386
Range	9
Minimum	1
Maximum	10
Sum	2761
Count	499
Confidence Level(95.0%)	0.258034

Types of mean

1)Weighted mean:

A weighted average is a type of mean that gives differing importance to the values in a dataset. In contrast, the regular average, or arithmetic mean, gives equal weight to all observations. The weighted average is also known as the weighted mean,

Weighted Mean Formula

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots + w_n}$$

where, **w** is the number of occurrences of x (weight)
x is the repeating value
$\bar{x}$ is the weighted mean
For example,

When Value of X = {1,2,3,4,5},
the Weights to the X values = {5,20,40,80,100}

Weighted mean= (W1X1+W2X2+W3X3+...WnXn)/(W1+W2+W3+...Wn)

Weighted mean=(5+40+120+320+500)/(5+20+40+80+100)

Weighted mean=(985)/(245)

Weighted mean=4.020408163265306

Weighted mean(rounded)=4.0204

Weighted mean in Excel :

Harmonic mean

The harmonic mean has uses in finance and technical analysis of markets, among others.

The harmonic mean is a type of numerical average. It is calculated by dividing the number of observations, or entries in the series, by the reciprocal of each number in the series. Thus, the harmonic mean is the reciprocal of the arithmetic mean of the reciprocals.

Q) What is HARMEAN Function?

A) HARMEAN Function is also a statistical function. The HARMEAN Function in excel calculates a **mean** which reduces the **impact of outliers** and returns the **harmonic mean** for the given data. The HARMONIC Mean is the **reciprocal of the arithmetic mean**.

In financial analysis, the HARMEAN Function calculates the **average** for **financial multiples**.

Syntax
=HARMEAN(number 1, [number 2],..)
Arguments:

- **=** – built-in function.
- **HARMEAN()** – function name.
- **number 1** – first value in given data.
- **number 2** – (optional) second value.

There can be up to 255 arguments to calculate the mean.

The HARMONIC Mean is also one among the three **Pythagoras mean** (*one is arithmetic mean and the other one is **geometric mean***). The HARMONIC Function is always **less than** the geometric mean and arithmetic mean. It is the measure of **ratio** and **rates**.

How to use HARMEAN Function in Excel ?

HARMONIC Function calculates ***Harmonic mean*** *which is the* ***reciprocal of the arithmetic mean****.*

Example:

STEP 1:
Open the workbook in Microsoft Excel.
STEP 2:
Enter the data in the workbook.

In this example, we entered random values to calculate **HARMEAN Function**.

STEP 3:

In the new cell, give the formula. First always start with ' =' for every functions, HARMEAN function name, followed by the open parenthesis, the arguments of the syntax.

The values are in the cell **A3:A7**.

STEP 4:

Press enter to get the results.

You get **79.86037** for the given data.

(a) H.M. for Ungrouped data

Let $x1, x2, \ldots, xn$ be the n observations then the harmonic mean is defined as

$$\text{H. M.} = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

Merits of H.M:

- It is rigidly defined
- It is based on all the observations of the series
- It is suitable in case of series having wide dispersion

Demits of H.M

- The harmonic mean is greatly affected by the values of the extreme items
- It cannot be able to calculate if any of the items is zero

Geometric mean:

The **Geometric mean** is an average that multiplies all values and finds a root of the number.

For a dataset with *n* numbers, you find the *n*th root of their product. You can use this descriptive statistic to summarize your data.

The geometric mean is an alternative to the arithmetic mean, which is often referred to simply as "the mean." While the arithmetic mean is based on adding values, the geometric mean multiplies values

Arithmetic Mean(average)

$$AM = \frac{a_1 + a_2 + a_3 + \ldots + a_n}{n}$$

Geometric Mean

$$GM = \sqrt[n]{a_1 a_2 a_3 \ldots a_n}$$

*e.g.*Find the AM and GM of 4 and 25

$$AM = \frac{25+4}{2} = \frac{29}{2}$$

$$GM = \sqrt{25 \times 4} = \sqrt{100} = 10$$

Using the geometric mean allows analysts to calculate the return on an investment that gets paid interest on interest. This is one reason portfolio managers advise clients to reinvest dividends and earnings.

The geometric mean is also used for present value and future value cash flow formulas.

The Excel **GEOMEAN** function returns the geometric mean for a set of numeric values. Geometric mean can be used to calculate average rate of return with variable rates.

The Excel Geomean function calculates the geometric mean of a supplied set of values.

The syntax of the function is:

GEOMEAN(number1, [number2], ...)

Where the number arguments are one or more positive numeric values (or arrays of numeric values), for which you want to calculate the geometric mean.

STEPS FOR CALCULATION OF GEOMETRIC MEAN IN EXCEL
(REFER LAB RECORD FOR PROCEDURE)

Advantages of Geometric mean

- A geometric mean is based on all observations.
- It is rigidly define
- The fluctuations of the observations do not effect the GM
- It gives more weight to the small items.

Disadvantages of Geometric mean

- It cannot be calculated if there is negative numbers.
- It cannot be found by inspection.
- It becomes zero if one of the observations is zero. of the observations is zero.

(a) G.M. For Ungrouped data

The Geometric Mean (G.M.) of a set of n observations is the nth root of their product. If $x1, x2, \ldots, xn$ are n observations then

$$\text{G. M.} = \sqrt[n]{x_1, x_2 \ldots x_n} = (x_1 . x_2 \ldots x_n)^{\frac{1}{n}}$$

Taking the nth root of a number is difficult. Thus, the computation is done as under

Statistic	*Excel Function*
Arithmetic Mean	AVERAGE(R1)
Median	MEDIAN(R1)
Mode	MODE.SNGL(R1), MODE.MULT(R1)
Geometric Mean	GEOMEAN(R1)
Harmonic Mean	HARMEAN(R1)

Empirical Relationship among mean, median and mode

A frequency distribution in which the values of arithmetic mean, median and mode coincide is known of symmetrical distribution, when the values of mean, median and mode are not equal the distribution is known as asymmetrical or skewed. In moderately skewed asymmetrical distributions a very important relationship exists among arithmetic mean, median and mode.

Karl Pearson has expressed this relationship as follows

Mode = 3 Median – 2 Arithmetic Mean

Mode = 3 Median – 2 Arithmetic Mean

Measures of Variability

or

Q) Determine core aspects of measurement and variability

A) **The concept of variability is a critical issue in the behavioral sciences, where research frequently examines differences in such things as characteristics, attitudes, and cognitive abilities.** Ultimately, the primary goal of a science such as psychology is to describe, understand, explain, and predict variability.
Researchers have developed statistics designed to measure variability. A measure of variability is a descript - tive statistic of the amount of differences in a set of data for a variable. **The purpose of measures of variability is to numerically represent a set of data based on how the scores differ or vary from each other.** Similar to mea- sures of central tendency, there are multiple measures of variability.

Four measures of Variability:

- The Range,
- The interquartile range,
- The variance, and
- The standard deviation.

Range

The range is the easiest dispersion of data or measure of variability. The range can measure by subtracting the lowest value from the massive Number. **The wide range indicates high variability, and the small range specifies low variability in the distribution**. To calculate a range, prepare all the values in ascending order, then subtract the lowest value from the highest value.

.
Mathematical formula :

Range = Highest_value – Lowest_value

Range is the difference between minimum and maximum value in a dataset.

Suppose you have a dataset of student scores and the range would be the spread of values in that data.

In this dataset, the minimum score is 15 and the maximum score is 98, so the spread would be 73.

Range = 98 – 15

Range simply refers to the difference between the maximum and the minimum value of the given dataset.

Excel Functions Used for Range

Excel lets you use different functions to make all kinds of calculations depending on the data available in the cells. Suppose, you have a dataset of numeric in cells A1 to A10. You can easily find out the highest number in the cells. Put the formula to use a function in a blank cell.

Below are some useful functions used in Excel related to the range:

MAX: Using this function, you can return the maximum value of a range. It will be as =MAX(A1:A10).

MIN: Using this function, you can return the minimum value of a range. It will be as =MIN(A1:A10).

SMALL: Using this function can return the smallest value in a given range and you can write it as =SMALL(A1:A10).

LARGE: Using this function can return the largest value in a given range and you can write it as =LARGE(A1:A10).

With Excel, you can calculate a range using the maximum and the minimum value from a given range.

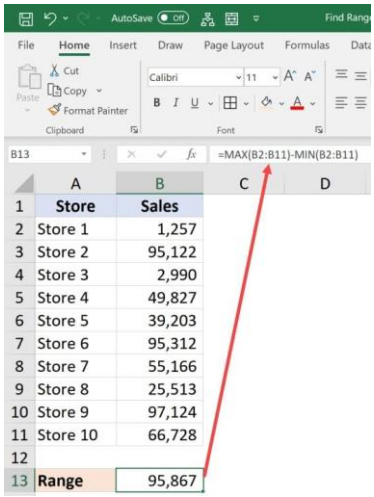
For instance, in a dataset, you need to find the range for the data in column B.

How to Calculate Range in Excel with the MAX and MIN Functions

Below are some easy-to-follow steps to find the range in Excel with the help of SMALL and LARGE functions:

- Add data in cells A1 to A10.
- Put the formula =**MIN**(A1:A10) in cell B1.
- Now, add the formula in cell B2 =**MAX**(A1:A10).
- Subtract the two formulas by adding =B2-B1 in cell C1.
- Now, you got the range of data.

	A	B
1	**Store**	**Sales**
2	Store 1	1,257
3	Store 2	95,122
4	Store 3	2,990
5	Store 4	49,827
6	Store 5	39,203
7	Store 6	95,312
8	Store 7	55,166
9	Store 8	25,513
10	Store 9	97,124
11	Store 10	66,728



How to Calculate Range in Excel with the SMALL and LARGE Functions

Below are some easy-to-follow steps to find the range in Excel with the help of SMALL and LARGE functions:

- Add data in cells A1 to A10.
- Put the formula =SMALL(A1:A10) in cell B1.
- Now, add the formula in cell B2 =LARGE(A1:A10).
- Subtract the two formulas by adding =B2-B1 in cell C1.
- Now, you got the range of data.

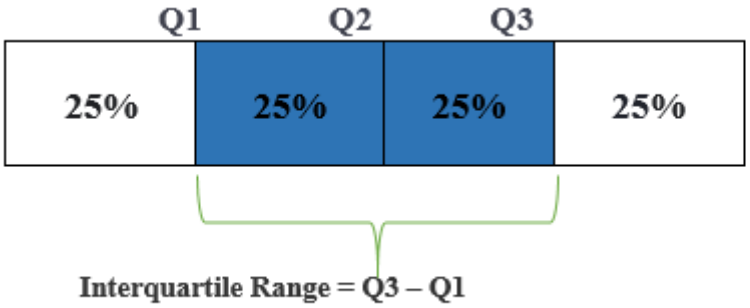
Advantages of Range :

- Range as a measure is easy to understand.
- It is simple to calculate.
- Range is widely used in some statistical series like change in interest rate, change in production etc.

Quartiles

Interquartile Range (IQR)

IQR is a range (*the boundary between the first and second quartile*) **and Q3** (**the boundary between the third and fourth quartile**). **IQR is preferred over a range as***, like a range, IQR does not influence by outliers***.** IQR is used to **measure variability** by splitting a data set into four equal quartiles.



Q_1 = [(n+1)/4]th item

Q_2 = [(n+1)/2]th item

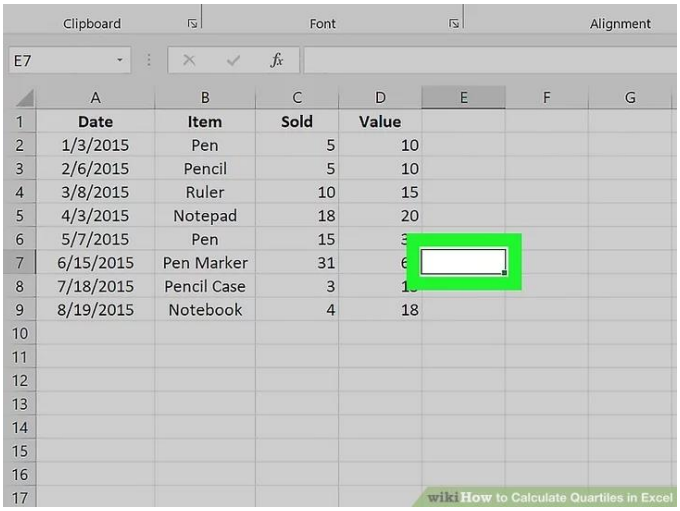
Q_3 = [3(n+1)/4]th item

A quartile is the name for a percentage of data in four parts, or quarters, which is especially helpful for marketing, sales, and teachers scoring tests.

The **QUARTILE** Function[1] is categorized under Excel Statistical functions. The function will return the quartile of a given data set.

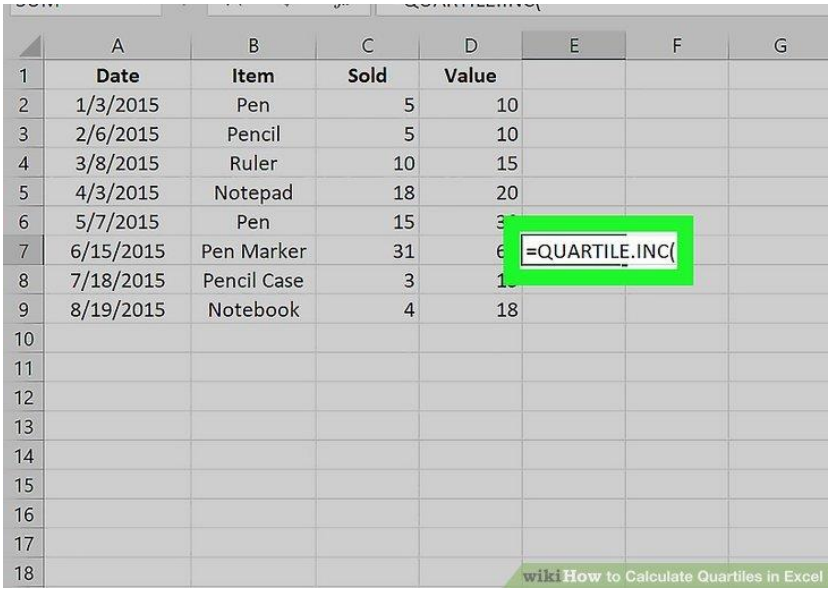
As a financial analyst, QUARTILE can be used to find out, for example, a specific percentage of incomes in a population. The function is very useful in revenue analysis.

- **Open your project in Excel.** If you're in Excel, you can go to **File > Open** or you can right-click the file in your file browser.



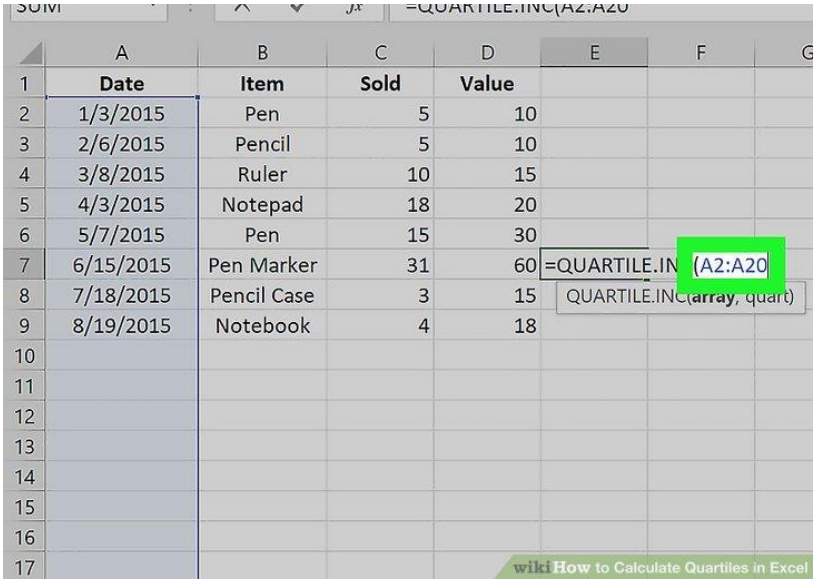
	A	B	C	D	E	F	G
1	Date	Item	Sold	Value			
2	1/3/2015	Pen	5	10			
3	2/6/2015	Pencil	5	10			
4	3/8/2015	Ruler	10	15			
5	4/3/2015	Notepad	18	20			
6	5/7/2015	Pen	15	3			
7	6/15/2015	Pen Marker	31	6			
8	7/18/2015	Pencil Case	3	1			
9	8/19/2015	Notebook	4	18			
10							
11							
12							
13							
14							
15							
16							
17							

- **Select an empty cell where you want to display your quartile information.** This can be anywhere on your spreadsheet.
  For example, you can select cell E7 even if all your data is located in cells A2-A20.

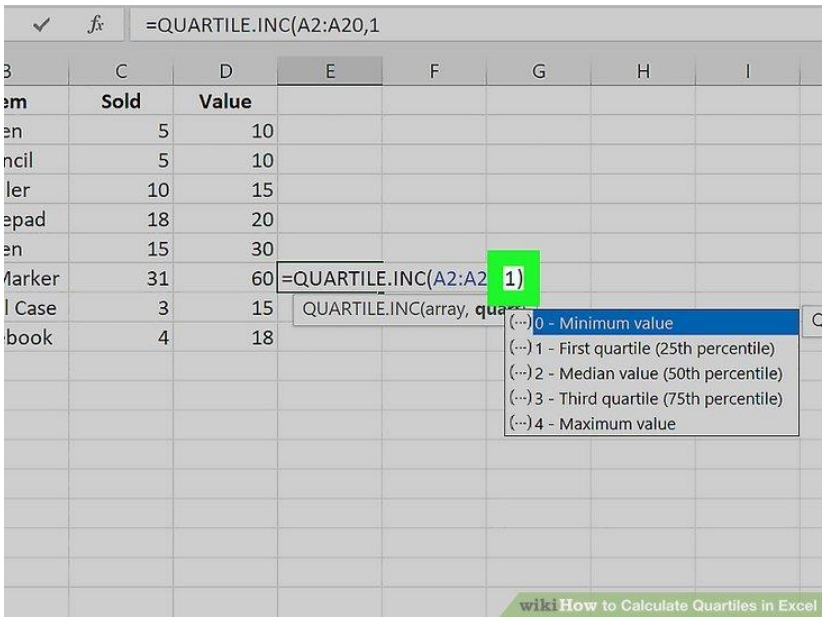


	A	B	C	D	E	F	G
1	Date	Item	Sold	Value			
2	1/3/2015	Pen	5	10			
3	2/6/2015	Pencil	5	10			
4	3/8/2015	Ruler	10	15			
5	4/3/2015	Notepad	18	20			
6	5/7/2015	Pen	15	3			
7	6/15/2015	Pen Marker	31	6	=QUARTILE.INC(		
8	7/18/2015	Pencil Case	3	1			
9	8/19/2015	Notebook	4	18			
10							
11							
12							
13							
14							
15							
16							
17							
18							

- **Enter the quartile function:** `=QUARTILE.INC(`. INC stands for "Inclusive," which will give you results that include 0+100.



- **Select the cells containing your data.** You can drag your cursor to select the entire range or you can select the first cell then press **CTRL + SHIFT + Down arrow**.



- After you've selected the data set, you'll see it entered into your formula. It'll look something like *"=QUARTILE.INC(A2:A20"*. Don't add the closing parentheses because you'll need to add more information to the function.

- **Enter ",1)" to finish the formula.** The number after the data range can represent either Q1, Q2, Q3, or Q4, so you can use any number 1-4 in the function instead of 1.
- The function **QUARTILE.INC(A2:A20,1)** will show you the first quartile (or 25th percentile) of your data set.

- **Press ↵ Enter (Windows) or ⏎ Return (Mac).** The cell you have selected will display the quartile function result. You can repeat this process using the other quartile function to see the differences.

Merits:

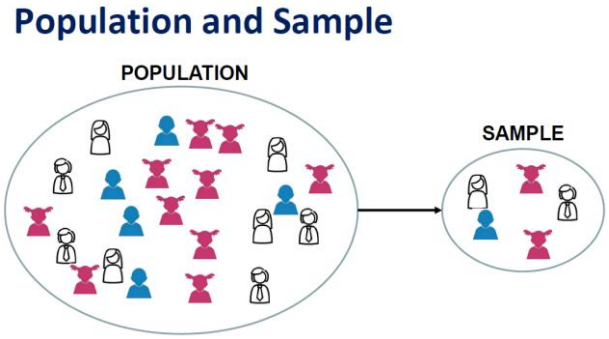
- It is not affected by the extreme (highest and lowest) values in the data set.
- It is an appropriate measure of variation for a data set summarized in open-ended class intervals.
- It is a positional measure of variation; therefore, it is useful in the cases of erratic or highly skewed distributions.

SAMPLE AND POPULATION IN STATISTICS

A population is a collection that consists of all possible data values and items within the field of study.

A population refers to the whole number of items or the entire group of people that are of interest in the statistical study.

A sample is a *subset* and a small portion of the population – a small part of all the possible data values that are part of the specified field of study.



Population and Sample Examples

1. All the students in the school are the population and the students of class 10 are the sample.
2. Patients in the hospital are the population and the old age patients are the sample.

Variance

Variance is the measure of variability of a data set that indicates how far different values are spread. Mathematically, it is defined as the average of the squared differences from the mean.

Mathematical Formula

Population	Sample
$\sigma^2 = \frac{\Sigma(x_i-\mu)^2}{n}$	$S^2 = \frac{\Sigma(x_i-\bar{x})^2}{n-1}$
μ - Population Average xi - Individual Population Value n - Total Number of Population σ^2 - Variance of Population	x - Sample Average x_i - Individual Population Value n - Total Number of Sample S^2 - Variance of Sample

To better understand what you are actually calculating with the variance, please consider this simple example.

In truth, variance just gives you a very general idea of the dispersion of the data set. A value of 0 means there is no variability, i.e. all the numbers in the data set are the same. The bigger the number, the more spread out the data.

How to calculate variance in Excel

There are 6 built-in functions to do variance in Excel:

1) VAR, 2)VAR.S 3)VARP 4)VAR.P 5)VARA and 6)VARPA

How to calculate sample variance in Excel

A **sample** is a set of data extracted from the entire population. And the variance calculated from a sample is called **sample variance**.

There are 3 functions to find sample variance in Excel:

VAR( ) , VARA( ), VAR.S( ), VAR.P ( )

VAR function in Excel

It is the oldest Excel function to estimate variance based on a sample. The VAR function is available in all versions of Excel 2000 to 2019.

VAR(number1, [number2], ...)

VAR.S function in Excel

It is the modern counterpart of the Excel VAR function. Use the VAR.S function to find sample variance in Excel 2010 and later.

VAR.S(number1, [number2], ...)

VARP function in excel

The Excel VARP function returns the variance of a population based on the entire set of numbers

VARP(number1, [number2], ...)

VAR.P function in Excel

It is an improved version of the VARP function available in Excel 2010 and later.

VAR.P(number1, [number2], ...)

VARA function in Excel

The Excel VARA function returns a sample variance based on a set of numbers, text, and logical values as shown in this table.

VARA(value1, [value2], ...)

Sample Variance formula in Excel

When working with a numeric set of data you can use any of the above functions to calculate sample variance in Excel.

As an example, let's find the variance of a sample consisting of 6 items (B2:B7). For this, you can use one of the below formulas:

=VAR(B2:B7)

=VAR.S(B2:B7)

=VARA(B2:B7)

As shown in the screenshot, all the formulas return the same result (rounded to 2 decimal places):

	A	B	C	D	E	F	G
1	Student	Score		Sample variance			
2	Daniela	85		VAR	34.97	=VAR(B2:B7)	
3	Tommy	79		VAR.S	34.97	=VAR.S(B2:B7)	
4	Edward	90		VARA	34.97	=VARA(B2:B7)	
5	Julia	88					
6	Timothy	88					
7	Peter	75					

How to calculate Population Variance in Excel

Population is all members of a given group, i.e. all observations in the field of study. **Population variance** describes how data points in the entire population are spread out.

VARP function in Excel

The Excel VARP function returns the variance of a population based on the entire set of numbers. It is available in all versions of Excel 2000 to 2019.

VARP(number1, [number2], ...)

VAR.P function in Excel

It is an improved version of the VARP function available in Excel 2010 and later.

VAR.P(number1, [number2], ...)

VARPA function in Excel

The VARPA function calculates the variance of a population based on the entire set of numbers, text, and logical values. It is available in all version of Excel 2000 through 2019.

VARA(value1, [value2], ...)

Population Variance formula in Excel

In the sample var calculation example, we found a variance of 5 exam scores assuming those scores were a selection from a bigger group of students. If you collect data on all the students in the group, that data will represent the entire population, and you will calculate a population variance by using the above functions.

Let's say, we have the exam scores of a group of 10 students (B2:B11). The scores constitute the entire population, so we will do variance with these formulas:

=VARP(B2:B11)

=VAR.P(B2:B11)

=VARPA(B2:B11)

And all the formulas will return the identical result:

	A	B	C	D	E	F	G
1	**Student**	**Score**		**Population variance**			
2	Daniela	85		VARP	36.41	=VARP(B2:B11)	
3	Tommy	79		VAR.P	36.41	=VAR.P(B2:B11)	
4	Edward	90		VARPA	36.41	=VARPA(B2:B11)	
5	Julia	88					
6	Timothy	88					
7	Peter	75					
8	Neal	92					
9	Sally	74					
10	Mike	83					
11	Adam	89					

Advantages of Variance:

1) Uses all of all data values.
2) Good measure
3) Used when data are Fairly symmetrical.

Standard Deviation

The **standard deviation** is a measure that indicates how much the values of the set of data deviate (spread out) from the mean. To put it differently, the standard deviation shows whether your data is close to the mean or fluctuates a lot.

The purpose of the standard deviation is to help you understand if the mean really returns a "typical" data. The closer the standard deviation is to zero, the lower the data variability and the more reliable the mean is. The standard deviation equal to 0 indicates that every value in the dataset is exactly equal to the mean. The higher the standard deviation, the more variation there is in the data and the less accurate the mean is.

Mathematical formula :

Population standard deviation

When you have collected data from every member of the population that you're interested in, you can get an exact value for population standard deviation.

The **population standard deviation** formula looks like this:

Formula	Explanation
$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$	• σ = population standard deviation • $\sum$ = sum of… • X = each value • μ = population mean • N = number of values in the population

Sample standard deviation

When you collect data from a sample, the sample standard deviation is used to make estimates or inferences about the population standard deviation.

The **sample standard deviation** formula looks like this:

Formula	Explanation
$s = \sqrt{\frac{\sum (X - \bar{x})^2}{n - 1}}$	• s = sample standard deviation • $\sum$ = sum of… • X = each value • $\bar{x}$ = sample mean • n = number of values in the sample

Excel makes calculating standard deviation more manageable. But first, it's important to understand the six standard deviation formulas in Excel.

1. To calculate the sample standard deviation, use formulas in this category: STDEV.S,
2. To calculate the standard deviation for an entire population, use formulas in this category: STDEV.P

Calculating standard deviation of a sample and population

Depending on the nature of your data, use one of the following formulas:

- To calculate standard deviation based on the entire **population**, i.e. the full list of values (B2:B50 in this example), use the STDEV.P function:

  =STDEV.P(B2:B50)

- To find standard deviation based on a **sample** that constitutes a part, or subset, of the population (B2:B10 in this example), use the STDEV.S function:

  =STDEV.S(B2:B10)

As you can see in the screenshot below, the formulas return slightly different numbers (the smaller a sample, the bigger a difference):

	A	B	C	D	E	F	G
1	**Name**	**Score**		**Population SD**	**15.78**	=STDEV.P(B2:B50)	
2	Neal	85		**Sample SD**	**14.44**	=STDEV.S(B2:B10)	
3	Alex	76					
4	Dora	73					
5	Tim	80					
6	Laura	72					
7	Kate	84					
8	Peter	79					
9	Isabella	120					
10	Lucas	85					
11	Ethan	73					
12	Olivia	77					
13	Ava	83					
14	Mason	108					

Advantages of Standard Deviation :

- The standard deviation uses all observations in a dataset in its calculation.
- The standard deviation is easy to interpret.

Disadvantages of standard deviation are :

- It doesn't give you the full range of the data.
- It can be hard to calculate.

Permutations and Combinations :

Permutation	Combination
In Permutation order of arrangement is important. For example, AB and BA are different combinations.	In Combination order of arrangement is not important. For example, AB and BA are the same combinations.
A permutation is used when different kinds of things are to be sorted or arranged.	Combinations are used when the same kind of things are to be sorted.
Permutation of two things out of three given things a, b, c is ab, ba, bc, cb, ac, ca.	the combination of two things from three given things a, b, c is ab, bc, ca.
Formula for permuation is: $^{n}\mathbf{P_r} = \mathbf{n!/(n-r)!}$	The formula for Combination is: $^{n}\mathbf{C_r} = \mathbf{n!/\{r! \times (n-r)!\}}$

Example 3: In how many ways can 5 different books be arranged on a shelf?

Solution:
This is a permutation problem because the order of the books matters.

Using the permutation formula, we get:

$^{5}P_{5} = 5! / (5-5)! = 5! / 0! = 5 \times 4 \times 3 \times 2 \times 1 = 120$
Therefore, there are 120 ways to arrange 5 different books on a shelf.

Example 5: A committee of 5 members is to be formed from a group of 10 people. In how many ways can this be done?

Solution:
This is a combination problem because the order of the members doesn't matter.

Using the combination formula, we get:

$^{10}C_{5} = 10! / (5! \times (10-5)!) = 10! / (5! \times 5!)$
$\Rightarrow {}^{10}C_{5} = (10 \times 9 \times 8 \times 7 \times 6) / (5 \times 4 \times 3 \times 2 \times 1) = 252$
Therefore, there are 252 ways to form a committee of 5 members from a group of 10 people.

Q) Explain Analyze Permutation with Repetition

A) Permutation formula is used to find the number of ways an object can be arranged without taking the order into consideration. To recall, when objects or

symbols are arranged in different ways and order, it is known as permutation. Permutation can be done in two ways,

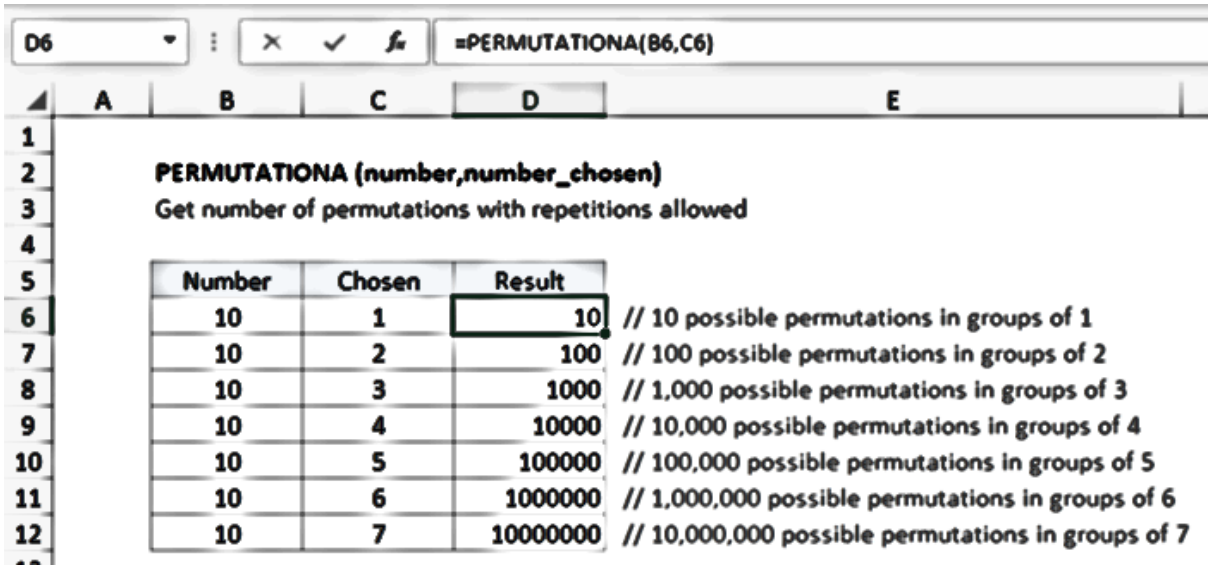
- **Permutation with repetition:** This method is used when we are asked to make different choices each time and with different objects.
- **Permutation without Repetition:** This method is used when we are asked to reduce 1 from the previous term for each time.

In Excel , we can calculate Permutation with Repetition using following function:

PERMUTATIONA ()

- The Excel PERMUTATIONA function returns the number of permutations (combinations where order is significant) for a given number of items.
- The PERMUTATIONA function *allows* repetitions.

Syntax: =PERMUTATIONA(number, number_chosen)



Example

To use PERMUTATIONA, specify the total number of items and "*number_chosen*", which represents the number of items in each combination

For -example, to calculate 3-number permutations for the numbers 0-9, there are 10 numbers and 3 chosen, so the formula is:

==PERMUTATIONA(10,3) turns 1000

Calculation of permutation without repetition using following function:

PERMUT()

The Excel PERMUT function returns the number of permutations (combinations where order is significant) for a given number of items. The PERMUT function *does not* allow repetitions.

Syntax

=PERMUT(number, number_chosen)

The PERMUT function returns the number of permutations for a given number of items. A permutation is a combination where order matters.

To use PERMUT, specify the total number of items and "*number_chosen*", which represents the number of items in each combination.

For example, to calculate 3-number permutations for the numbers 0-9, there are 10 numbers and 3 chosen, so the formula is:

=PERMUT(10,3) // returns 720

Q) Explain Analyze Combinations without Repetition.

A) It is a very useful and important technique that can be used to determine the number of ways that a set of objects can be combined or arranged when order does not matter.

can be done in two ways,

- **Combinations with repetition:** This is when the elements of a set can be repeated, to clarify this type **.**
- **Combinations without Repetition:** It is when the elements of a set cannot be repeated, for example: in a company where there work 20 people they take a decision of forming a directive composed by 3 people, in this case we would have a combination without repetition, because a person cannot be chosen twice.

In excel , order to find the possible number of **combinations with out repetition**

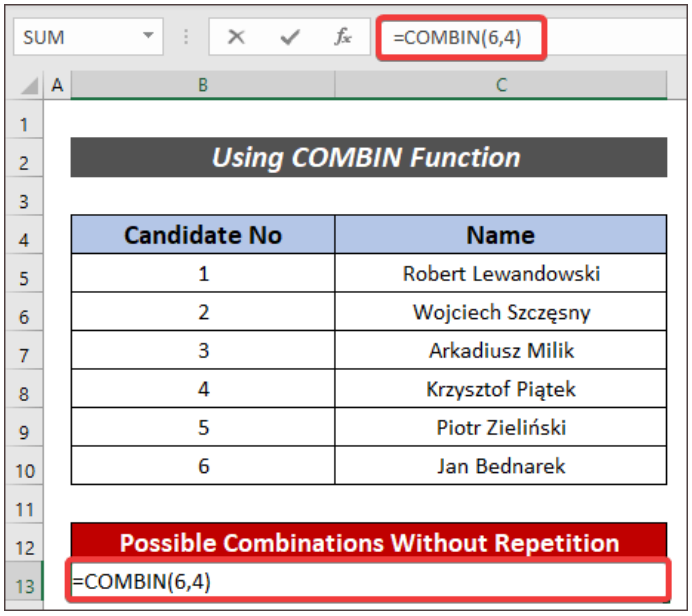
the **COMBIN() function** can be used.

The **COMBIN Function**[1] is an Excel Math and Trigonometry function. The function will calculate the number of combinations without repetitions for a given number of items. It was introduced in MS Excel 2000.

Syntax:

COMBIN(number, number_chosen)
Returns the number of combinations for a given number of items.
number The total number of items.

number_chosen The number of items in each combination.



Combination with Repetitions

The important fact here is that it just gives the number of combinations, but the actual combinations.

The COMBINA Function[1] is an Excel Math and Trigonometry function. It will calculate for a given number of items the number of combinations, along with repetitions. The function was introduced in MS Excel in 2013 and hence not available in earlier versions.

Formula

=COMBINA(number, number_chosen)

The COMBINA function uses the following arguments:

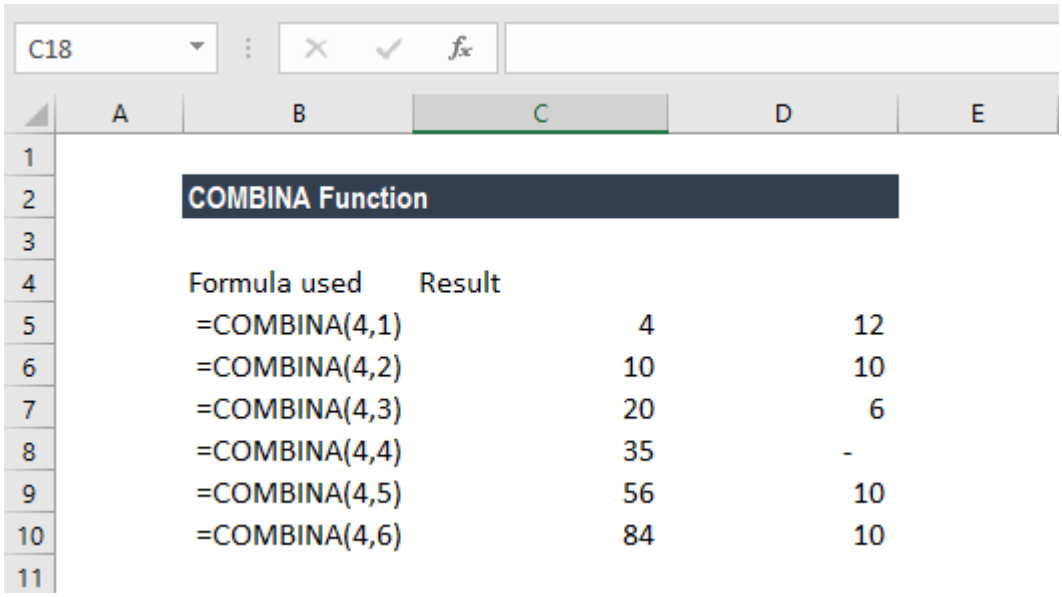
1. **Number** (required argument) – This should be either greater than or equal to 0. Also, it should be greater than or equal to number_chosen. Non-integer values, if entered for this parameter, are truncated.
2. **Number_chosen** (required argument) – This must be greater than or equal to 0. Non-integer values are truncated.

How to use COMBINA Function in Excel?

To understand the uses of the COMBINA function, let's consider a few examples:

Example 1

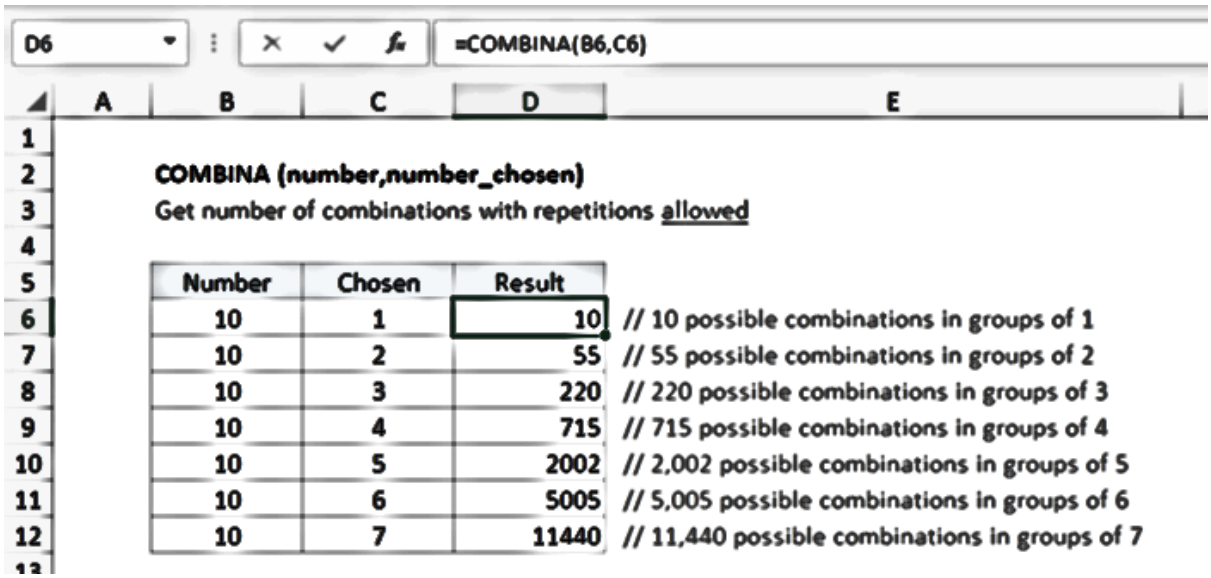
Suppose we are given six numbers: 1, 2, 3, 4, 5, 6. Now let's see how the COMBINA function calculates the number of combinations (with repetitions) for other numbers of objects taken from a set of 4.



	A	B	C	D	E
1					
2		COMBINA Function			
3					
4		Formula used	Result		
5		=COMBINA(4,1)	4	12	
6		=COMBINA(4,2)	10	10	
7		=COMBINA(4,3)	20	6	
8		=COMBINA(4,4)	35	-	
9		=COMBINA(4,5)	56	10	
10		=COMBINA(4,6)	84	10	
11					

For any six objects – a, b, c, d, e, f – there are 21 different combinations (with repetitions) of 2 objects. These are:

aa, ab, ac, ad, ae, af, bb, bc, bd, be, bf, cc, cd, ce, cf, dd, de, df, ee, ef, ff



D6 =COMBINA(B6,C6)

COMBINA (number,number_chosen)

Get number of combinations with repetitions allowed

Number	Chosen	Result	
10	1	10	// 10 possible combinations in groups of 1
10	2	55	// 55 possible combinations in groups of 2
10	3	220	// 220 possible combinations in groups of 3
10	4	715	// 715 possible combinations in groups of 4
10	5	2002	// 2,002 possible combinations in groups of 5
10	6	5005	// 5,005 possible combinations in groups of 6
10	7	11440	// 11,440 possible combinations in groups of 7

PROBABILITY FOR DATA ANALYTICS :
Basic Probability: Uses of probability - Differentiate between sample space, event, independent and dependent - Calculate probability - Probability and Ven Diagramming: Analyze “this” OR “that” diagram - Analyze “this” AND “that” diagram - Analyze exclusive diagram – Joint probability - Conditional probability - Calculating Probability: Calculate P using a contingency table - Calculate P from trees - Calculate Bayes’ theorem - Calculate the mean in terms of probabilities - Calculate the variance and standard deviation in terms of probabilities – Calculate conditional probability

Terms Related to Probability

1. **Experiment:** A trial or an operation conducted to produce an outcome is called an experiment.
2. **Trial:** A trial denotes doing a random experiment.
3. **Random Experiment:** An experiment with a known set of possible outcomes is called a random experiment if it cannot be predicted in advance which specific outcome will occur on a given execution of the experiment. Random experiments include things like coin tosses, die rolls, and card draws from a deck.
4. **Outcome:** The result of any random experiment is called an outcome. Suppose you tossed a coin and got head(H) as the upper surface. So, tossing a coin is a random experiment that results in an outcome ‘head’.
5. **Sample Space:** It is a set of all the possible outcomes for a random experiment. For example - Obtaining a head or a tail on the tossing of a coin is possible. Thus, S={H, T}: the Head and Tail are the sample space for tossing a coin. Similarly, on rolling a die, we can get either of the following numbers - 1, 2, 3, 4, 5, 6. Thus, S={1, 2, 3, 4, 5, 6} are the sample space for rolling a die.
6. **Equally Likely Outcomes:** Events with the same theoretical probability (or likelihood) of occurring are referred to as equally likely events. For example, the relative occurrences of Head and Tail on tossing a coin for a very large number of tosses are equal. So, Head and Tail are equally likely outcomes that make the tossing of a coin fair and unbiased if it's to decide between two options.
7. **Event:** In the case of a random experiment, an event is a set of possible outcomes at a specified condition. Example - On rolling of a die, 4 is not obtained. This event is the random experiment that is rolling of a die whose result is not 4. Thus, this event has 5 possible outcomes that are 1, 2, 3, 5, and 6. Suppose it's mentioned that the event F is equal to obtaining a black card while drawing a card from a deck of 52 cards. In this case, event F has 26 possible outcomes because there are 26 black cards, a total that is 13 spades and 13 clubs.

Q) What is Probability? List out uses of Probability.

A) Probability denotes the possibility of the outcome of any random event.

The meaning of this term is to check the extent to which any event is likely to happen.

For example, when we flip a coin in the air, what is the possibility of getting a head? The answer to this question is based on the number of possible outcomes. Here the possibility is either head or tail will be the outcome. So, the probability of a head to come as a result is 1/2.

Applications of Probability:

1. It plays a great role in making predictions.
2. It is used to solve scientific investigations as well as day-to-day problems.
3. It is very useful when one wants to predict the uncertainties of betting.and chances of success.
4. It is extensively used in business and economics.
5. It is the basis of the law of statistical regularity and law of inertia of a large number.

Uses of Probability

Following are the business situations where managers use probability to take a decision :

1) Investment Problem :
A company's manager has two different projects (with different initial costs) for investment. The decision taken by the manager is based on the choice that the outcome of which is dependent on the level of demand.

2) Introducing a New Product :
While introducing a new product there arises a problem in deciding about the introduction of the product. The decision maker is not sure whether the product would be acceptable or not. Hence after conducting test, marketing in three different regions the results being contradictory, the idea of introducing a new product should be dropped ? Thus, it is necessary to answer this question : what would be the chances of success of introducing a new product ?.

3) Stocking Decisions :

The demand of a perishable commodity is unknown to its dealer in advance. Thus if the commodity is not sold by the end of the day it will be spoiled. Therefore, he is not sure about the demand pattern and stock amount (that how much commodity is to be stored in advance).

4) The Individual Investor :
Maximisation of returns is the motive of investors who are engaged in buying and selling of equities. The behavior of equities/ security prices is uncertain and it depends upon various factors. In this situation, managers take decisions on the basis of their prediction about future prices of the securities. This prediction helps the investor to take a decision as to which.securities he/she should select for the investment.

Q) What is Sample Space and Event? Explain difference between Sample space and Event

Sample Space

In the study of probability, an **experiment** is a process or investigation from which results are observed or recorded.

An **outcome** is a possible result of an experiment.

A **sample space** is the set of all possible outcomes in the experiment. It is usually denoted by the letter *S*. Sample space can be written using the set notation, { }.

Sample Points - Each outcome in the sample space is known as a sample point. These are also called elements. For example, I throw an unbiased dice. I can say that there are a total of six sample points in the sample space.

Experiment 1: Tossing a coin
Possible outcomes are head or tail.
Sample space, S = {head, tail}

Sample points – head , tail

Experiment 2: Tossing a die

Possible outcomes are the numbers 1, 2, 3, 4, 5, and 6
Sample space, S = {1, 2, 3, 4, 5, 6}

Sample points =1,2,3,4,5,6

The three most common ways to find a sample space are:

1. To List All the Possible Outcomes.
2. Create a Tree-Diagram.
3. Use a Venn Diagram.

Q) What is an Event? list some Types of Events.

A) Event: -

In probability, the set of outcomes from an experiment is known as an Event. So say for example you conduct an experiment by tossing a coin. The outcome of this experiment is the coin landing 'heads' or 'tails'. These can be said to be the events connected with the experiment. So when the coin lands tails, an event can be said to have occurred.

Example 1 Tossing a coin. The sample space is S = {H, T}. E = {H} is an event

Example 2 Tossing a die. The sample space is S = {1, 2, 3, 4, 5, 6}. E = {2, 4, 6} is an event, which can be described in words as "the number is even".

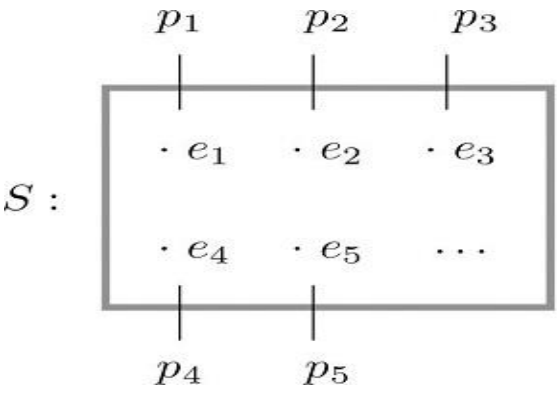
The probability of an event A is the sum of the probabilities of the individual outcomes of which it is composed. It is denoted P(A).

P(E) = Number of Favourable Outcomes/ Total Number of Outcomes

The following formula expresses the content of the definition of the probability of an event :

If an event E is E={e1, e2, ...,ek}= then P(E)=P(e1)+P(e2)+...+P(ek)

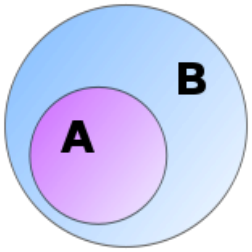
The following figure expresses the content of the definition of the probability of an event
:



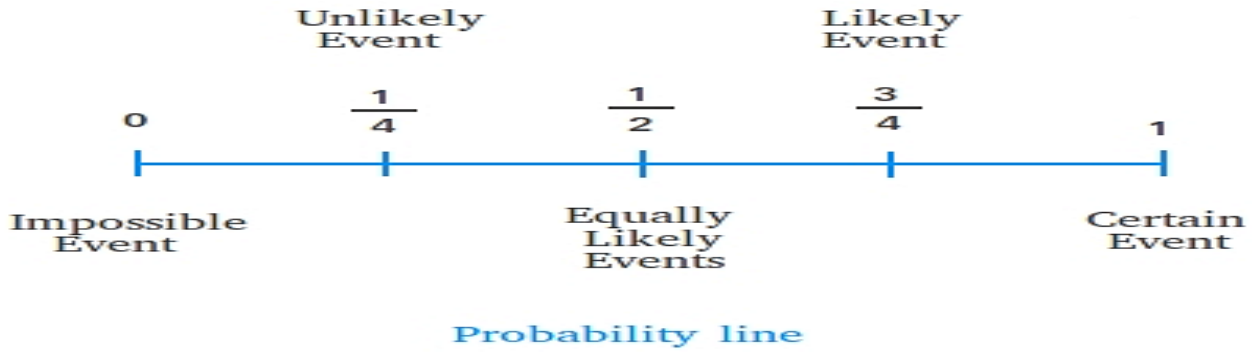
$$A: \quad \{e_1, e_2\}$$
$$B: \quad \{e_2, e_3, e_4\}$$

$$P(A) = p_1 + p_2$$
$$P(B) = p_2 + p_3 + p_4$$

Since the whole sample space S is an event that is certain to occur, the sum of the probabilities of all the outcomes must be the number 1.



So if you look at the Venn Diagram, B is the Sample Set which defines all the probable outcomes of an experiment. Subset A is the Event Space, so an event will be said to have occurred if it falls into the subset.



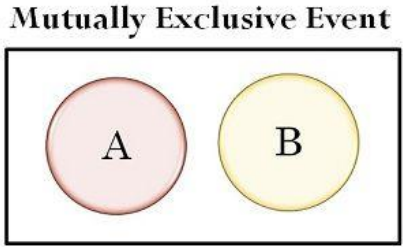
Probability line

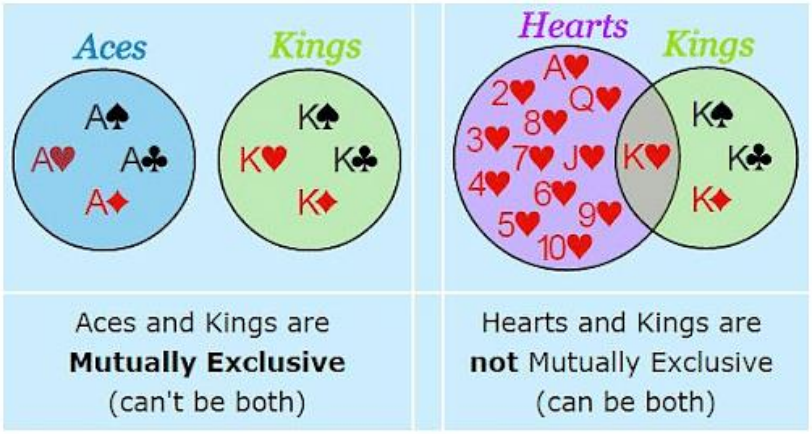
Types of Events in Probability:

1)Mutually Exclusive Events

Two events can be called mutually exclusive events when both of them cannot occur exactly at the same time. These types of events always have a different outcome.

If the occurrence of one event prevents the occurrence of another event, such events are mutually exclusive events. It implies that the two events don't have any common ground to bond.





Example of a mutually exclusive event is,

if S = {1, 2, 3, 4, 5, 6} and E1, E2 are two events such that E1 consists of numbers less than 3 and E2 consists of numbers greater than 4.

So, E1 = {1,2} and E2 = {5,6} .

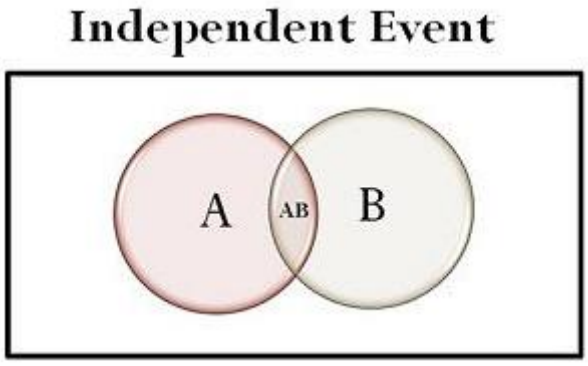
Then, E1 and E2 are mutually exclusive.

If A and B are two events, then
$P(A \cap B) = \emptyset$
$P(A \cap B) = 0$
$P(A \cup B) = P(A) + P(B)$

2) Independent Events

If the occurrence of an event remains uninfluenced by any other event, then we can call it an independent event. However, in probability, where the occurence of one event is affected by another is known as dependent events.

Two events are said to be independent, when the occurrence of one event cannot control the occurrence of other.



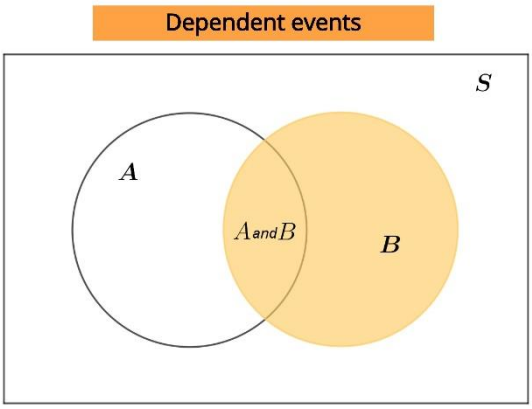
Example of an independent event is,

Putting on your favourite nail paint and failing a test, or winning a game of charades and drinking milk. These pairs of activities do not influence each other in any way and therefore are independent of each other.

Formula: $P(A \cap B) = P(A)\,P(B)$

Dependent Events:

Two events are said to be dependent if the outcome of one event affects the outcome of the other.



In probability, dependent events are usually real-life events and rely on another event to occur. For example, Sam scored well in his math test because he studied for it; the gym class had a football session because Adam got a football from home. If you look at these examples, then you will notice that one event is dependent on the other from happening.

Example of a dependent event is,
Getting into an accident because of rash driving, remaining hungry at night because you ended up cooking burnt food, getting yourself fined because you parked your vehicle in the middle of the road, can all be called as dependent events. One event's outcome leads to another.

Q) Write the differences between Dependent Event and Independent Event

A) There are two types of events in probability which are often classified as dependent or independent events.

The difference is given below in the table.

Dependent Events	Independent Events
1. The occurrence of one event affecting the probability of another event.	1. The occurrence of one event not affecting the probability of another event.
2.Examples include a power cut in case you don't pay your bill on time, winning the lottery after buying 10 lottery tickets (the more the tickets bought, the greater the chance of winning)	Examples include riding a bike and watching your favorite movie on a laptop
3. Formula can be written as: **$P(A \cap B) = = P(A) \times P(B \mid A)$**	3. Formula can be written as: **$P(A \cap B) = P(A) \times P(B)$**

Independent	Dependent
When two or more events happen, the occurrence of one event ***does not affect*** the probability of the second event.	When two or more events happen, the occurrence of one event ***does affect*** the probability of the second event.
Examples: Flipping a coin, tossing a dice, spinning a spinner.	Examples: Drawing an item from a bag then not replacing it.
Independent: the marble is replaced	Dependent: the marble is not replaced
An event is only independent if the number of outcomes stays the same.	An event is only dependent if the number of outcomes stays the same changes.

Example 1: What is the probability that a card taken from a standard deck, is an Ace?

Solution:

Total number of cards a standard pack contains = 52

Number of Ace cards in a deck of cards = 4

So, the number of favourable outcomes = 4

Now, by looking at the formula,

Probability of selecting an ace from a deck is,

P(Ace) = (Number of favourable outcomes) / (Total number of favourable outcomes)

P(Ace) = 4/52

= 1/13

So we can say that the probability of getting an ace is 1/13.

Example 2: Calculate the probability of getting an odd number if a dice is rolled.

Solution:

Sample space (S) = {1, 2, 3, 4, 5, 6}

n(S) = 6

Let "E" be the event of getting an odd number, E = {1, 3, 5}

n(E) = 3

So, the Probability of getting an odd number is:

P(E) = (Number of outcomes favorable)/(Total number of outcomes)

= n(E)/n(S)

= 3/6

= ½

Basic Probability Formulas

Let A and B are two events. The probability formulas are listed below:

All Probability Formulas List in Maths	
Probability Range	$0 \le P(A) \le 1$
Rule of Addition	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Rule of Complementary Events	$P(A') + P(A) = 1$
Disjoint Events	$P(A \cap B) = 0$
Independent Events	$P(A \cap B) = P(A) \cdot P(B)$
Conditional Probability	$P(A \mid B) = P(A \cap B) / P(B)$
Bayes Formula	$P(A \mid B) = P(B \mid A) \cdot P(A) / P(B)$
Joint probability	$P(A \cap B) = P(A) \cdot P(B)$

Q) What is Probability? Explain calculation probability in detail

OR

How we will do Probability calculation in excel?

A) The probability formula defines the likelihood of the happening of an event. It is the ratio of favourable outcomes to the total favourable outcomes.

How to Calculate the Probability Step by Step

You can use the following steps to calculate the probability of an event:

Step 1: Identify an event with one result.

Step 2: Identify the total number of results or outcomes and favourable outcomes that can occur.

Step 3: Divide the number of favourable outcomes by the total number of possible outcomes.

Formula

The most common formula used to determine the likelihood of an event is given below:

$$P(A) = \frac{Number\ of\ Favourable\ Outcome}{Total\ Number\ of\ Favourable\ Outcomes}$$

Or

$$\boldsymbol{P(A) = \frac{n(A)}{n(S)}}$$

where,

A) P(A) is the probability of an event 'A'.

B) n(A) is the number of favorable outcomes of an event 'A'.

C) n(S) is the total number of events occurring in a sample space.

D) Probability calculation can be done using function **PROB ()**

What is the PROB Function?

The PROB Function[1] is categorized under Excel Statistical functions. It will calculate the probability that is associated with a given range.

In financial analysis, the PROB function can be useful in estimating business losses. It can also be used by a risk manager to carry out financial probability analysis.

Formula

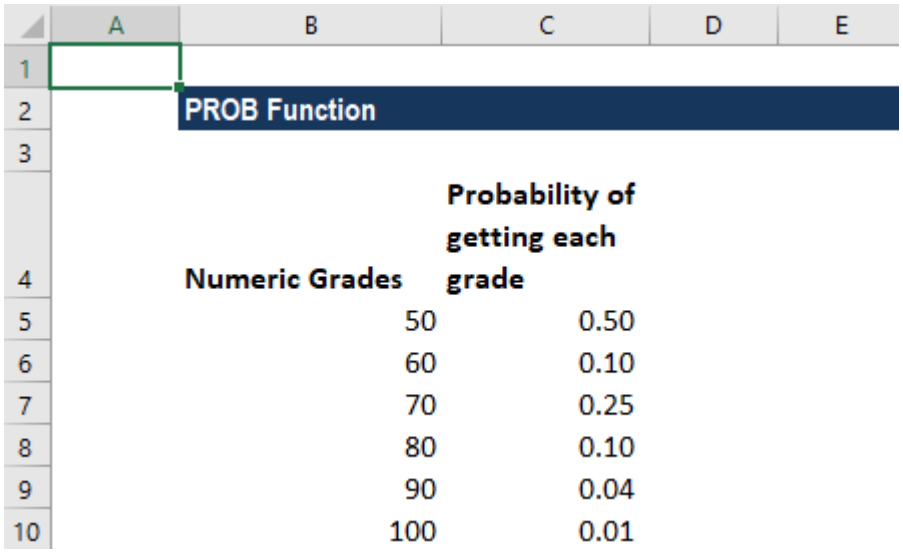
=PROB(x_range, prob_range, [lower_limit], [upper_limit])

The PROB function uses the following arguments:

1. **X_range** (required argument) – This is the range of numeric values of x with which there are associated probabilities.
2. **Prob_range** (required argument) – This is the set (array) of probabilities that is associated with values in x_range. The array must be of the same length as the x_range array and the values in prob_range must add up to 1.
3. **Lower_limit** (optional argument) – This is the lower boundary of the value for which we want a probability.
4. **Upper_limit** (optional argument) – This is upper boundary of the value for which you want a probability. When we omit this argument, the PROB function will simply return the probability associated with the value of the supplied lower_limit.

Example 1

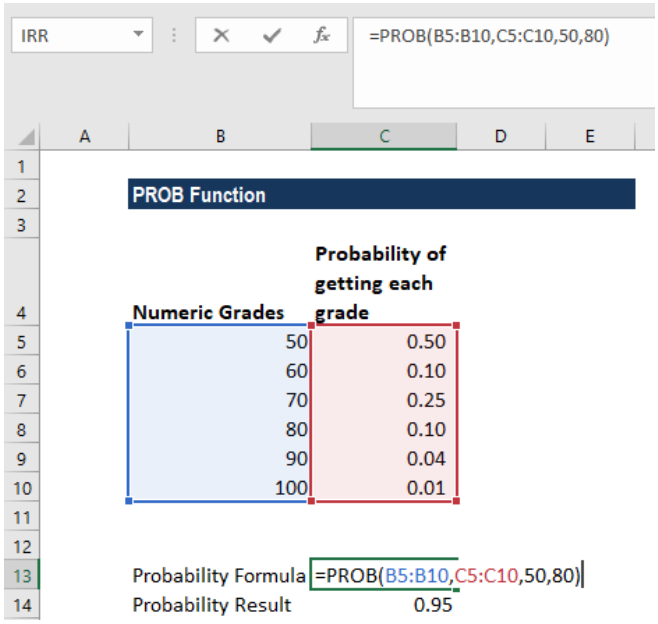
Suppose we are given the data below:



	A	B	C	D	E
1					
2		PROB Function			
3					
4		Numeric Grades	Probability of getting each grade		
5		50	0.50		
6		60	0.10		
7		70	0.25		
8		80	0.10		
9		90	0.04		
10		100	0.01		

The probability for the given range when the lower limit is set to 50 and the upper limit is 80 would be:

- Select any cell type probability
- Select another cell adjacent to probability cell and type = **prod(.....**
- Select B5 – B10 for X- range limit,
- Select C5 – C10 for probability range
- Enter lower limit 50, upper limit 80, press enter



Example 2:

	Product Sales	Probability
	20	0.23
	50	0.50
	80	0.18
	110	0.09
Lower limit	50	
Upper limit	80	
Probability		

To perform the calculation, we enter this formula in cell C11

=PROB(B4:B7,C4:C7,C9,C10)

Where B4:B7 is the range containing the values for product sales, C4:C7 contains the probabilities for each sales quantity, C9 is the lower limit of 50 while C10 is the upper limit of 80.

C11 =PROB(B4:B7,C4:C7,C9,C10)

	Product Sales	Probability
	20	0.23
	50	0.50
	80	0.18
	110	0.09
Lower limit	50	
Upper limit	80	
Probability	0.68	

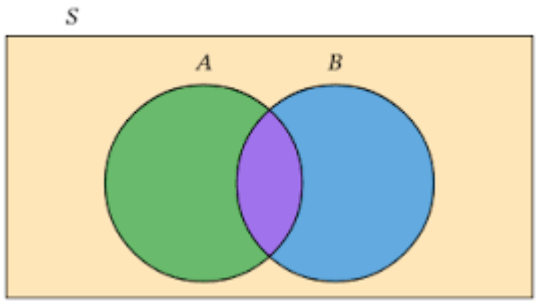
Probability with Venn Diagram

Q) What is Venn Diagram?

A) A Venn diagram is a visual tool used to show the relationships between various sets or events. Most commonly, a Venn diagram consists of circles that partially overlap, and these circles represent individual events.

In probability, a Venn diagram is a figure with one or more circles inside a rectangle that describes logical relations between events. The rectangle in a Venn diagram represents the sample space or the universal set, that is, the set of all possible outcomes. A circle inside the rectangle represents an event, that is, a subset of the sample space.

We consider the following Venn diagram involving two events, *A* and *B*.



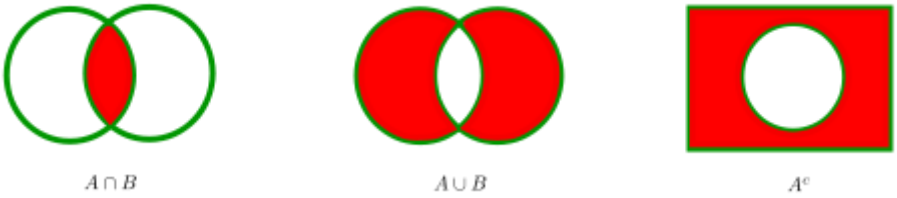
In the diagram above, we have two events A and B within the sample space (or universal set) S. Sometimes, the sample space is denoted by σ or ξ instead of S.

Coloured regions in this Venn diagram represent the following events:

Definition: Two-Event Venn Diagrams or "this" , "that"

Let A and B be events described in a Venn diagram. Then,

- the circles do not overlap if A and B are mutually exclusive events, that is, $A \cap B = \emptyset$;
- the circles overlap if $A \cap B \neq \emptyset$, in which case the intersection $A \cap B$ is represented by the overlapping region;
- the region outside both circles but within the rectangle represents the complement of the union of both events, that is, $A \cup B$ or, alternatively, $(A \cup B)'$.

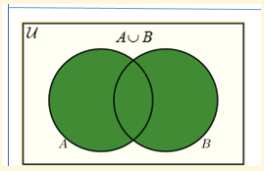
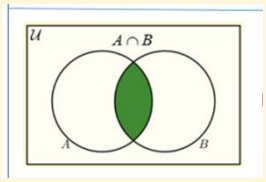
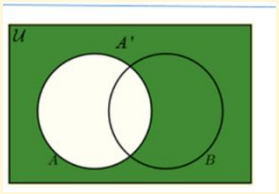


Within each divided region of a Venn diagram, we can add data in one of the following ways:

- the outcomes of the event,
- the number of outcomes in the event,
- the probability of the event.

Venn Diagram Symbols

There are more than 30 Venn diagram symbols. We will learn about the three most commonly used symbols in this section. They are listed below as:

Venn Diagram Symbols	Explanation	Venn Diagram
The union symbol - ∪	A ∪ B is read as A union B. Elements that belong to either set A or set B or both the sets. U is the universal set.	
The intersection symbol - ∩	A ∩ B is read as A intersection B. Elements that belong to both sets A and B. U is the universal set.	
The complement symbol - A^c or A'	A' is read as A complement. Elements that don't belong to set A. U is the universal set.	

Applications of Venn Diagram

There are several advantages to using Venn diagrams. Venn diagram is used to illustrate concepts and groups in many fields, including statistics, linguistics, logic, education, computer science, and business.

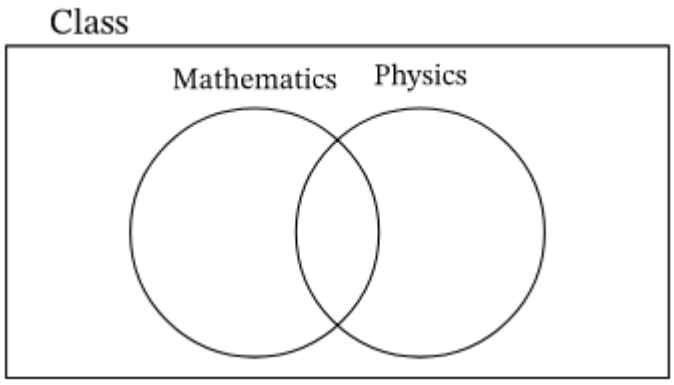
- We can visually organize information to see the relationship between sets of items, such as commonalities and differences, and to depict the relations for visual communication.
- We can compare two or more subjects and clearly see what they have in common versus what makes them different. This might be done for selecting an important product or service to buy.
- Mathematicians also use Venn diagrams in math to solve complex equations.
- We can use Venn diagrams to compare data sets and to find correlations.
- Venn diagrams can be used to reason through the logic behind statements or equations.

Example 1: Organizing Data using Venn Diagrams to Find Probabilities

A class contains 100 students; 70 of them like mathematics, 60 like physics, and 40 like both. If a student is chosen at random, using a Venn diagram, find the probability that they like mathematics but not physics.

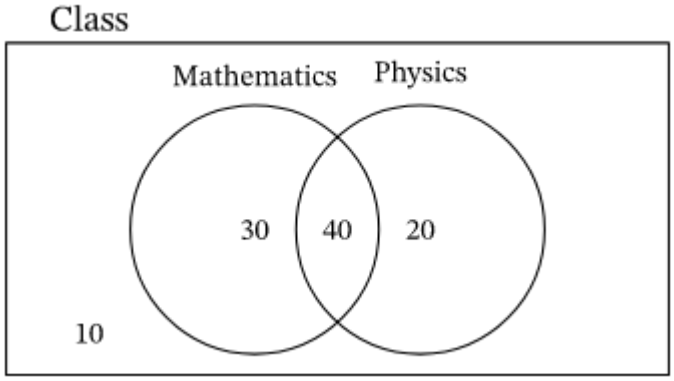
Answer

We begin by drawing an empty Venn diagram to represent this example.

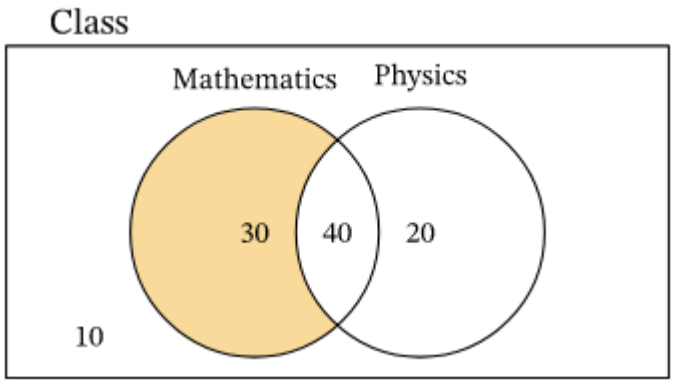


We know that the overlapping region in a Venn diagram represents the intersection of the events. We are given that 40 students belong to this intersection, since they like both mathematics and physics.

70 students like mathematics, and 40 of them also like physics. This tells us that the number of students who like mathematics only is 70−40=30.Similarly, we can deduce that 60−40=20 students like physics only. This leads to the following Venn diagram.



We note that the number 10 outside is to ensure that the sum of all values within the Venn diagram is equal to 100, since the class contains 100 students total. We are looking for the probability that a randomly chosen student likes mathematics but not physics. The region of the Venn diagram representing this event is highlighted below.



From the Venn diagram, we note that 30 students like mathematics but not physics. Since a student is randomly selected, we can obtain the probability of this event when dividing this number by the total number of students, which is 100.

Hence, the probability that a randomly selected student likes mathematics but not physics is 30/100=0.3.

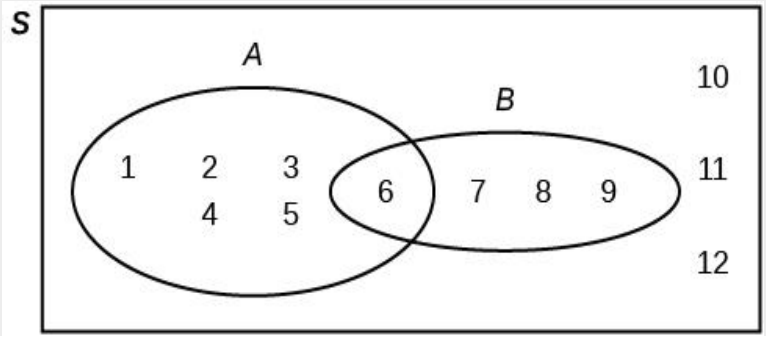
Example 2 :

Suppose an experiment has the outcomes 1, 2, 3, … , 12 where each outcome has an equal chance of occurring.

Let event A = {1, 2, 3, 4, 5, 6} and event B = {6, 7, 8, 9}.

Then A intersect B = $A \cap B = \{6\}$ and A union B = $A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

The Venn diagram is as follows:



Example 3:

Flip two fair coins.

Let A = tails on the first coin.

Let B = tails on the second coin.

Then $A = \{TT, TH\}$ and $B = \{TT, HT\}$.

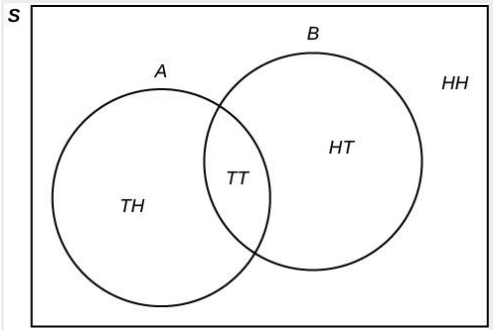
Therefore, A∩B = {TT}

A∪B = {TH, TT, HT}

The sample space when you flip two fair coins is $X = \{HH, HT, TH, TT\}$.

The outcome *HH* is in NEITHER *A* NOR *B*.

The Venn diagram is as follows:



For more examples : https://openstax.org/books/introductory-business-statistics/pages/3-5-venn-diagrams

Joint Probability

Q) EXPLAIN JOINT PROBABILITY.

A) Joint **Probability**

Joint probability is the likelihood of **more than one event** occurring at the **same time** P(A and B). The probability of event A and event B occurring together. Joint probability is the probability of event Y occurring at the same time that event X occurs. It is the probability of the intersection of two or more events written as **p(A ∩ B)** or **p(A,B)**

Conditions for Joint Probability

- One is that events X and Y must happen at the same time. **Example:** *Throwing two dice simultaneously*.
- The other is that events X and Y must be independent of each other. That means the outcome of event X does not influence the outcome of event Y.

  Example: *Rolling two Dice.*
- If the following conditions met, then **P(A∩B) = P(A) * P(B).**

Let's consider a simple example. A bag contains 10 blue and 10 red balls if we choose 1 red and 1 blue from the bag on a single take. What will be the joint probability of choosing 1 blue and 1 red?

Solution:

- Possible outcomes = (red, blue),(blue, red),(red, red), (blue, blue)=4
- Favorable outcomes = (red, blue) or (blue, red) = 1

Use the below-given data for the calculation.

Particulars	Value
Number of Red Balls	10
Number of Blue Balls	10
Probability of Choosing 1 Red and 1 Blue	1
Possible Outcomes	4

Probability of choosing the red ball

B7 =B5/B6

	A	B
2	**Particulars**	**Value**
3	Number of Red Balls	10
4	Number of Blue Balls	10
5	Probability of choosing 1 Red and 1 Blue	1
6	Possible Outcomes	4
7	**Probability of choosing Red Ball**	0.25
8		

- **P (a) = 1/4**
- **= 0.25**

Probability of choosing a blue ball

B8 =B5/B6

	A	B
2	**Particulars**	**Value**
3	Number of Red Balls	10
4	Number of Blue Balls	10
5	Probability of choosing 1 Red and 1 Blue	1
6	Possible Outcomes	4
7	Probability of choosing Red Ball	0.25
8	**Probability of choosing Blue Ball**	0.25
9		

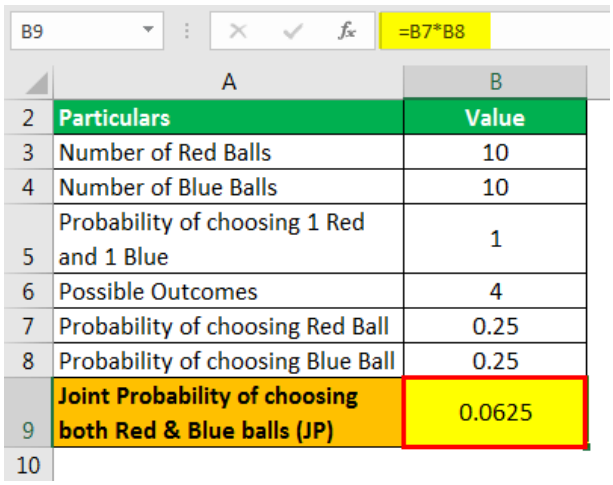
- P (b) = 1/4
- **= 0.25**

Joint Probability Formula (JP) = P (A)*P (B)

=B7*B8

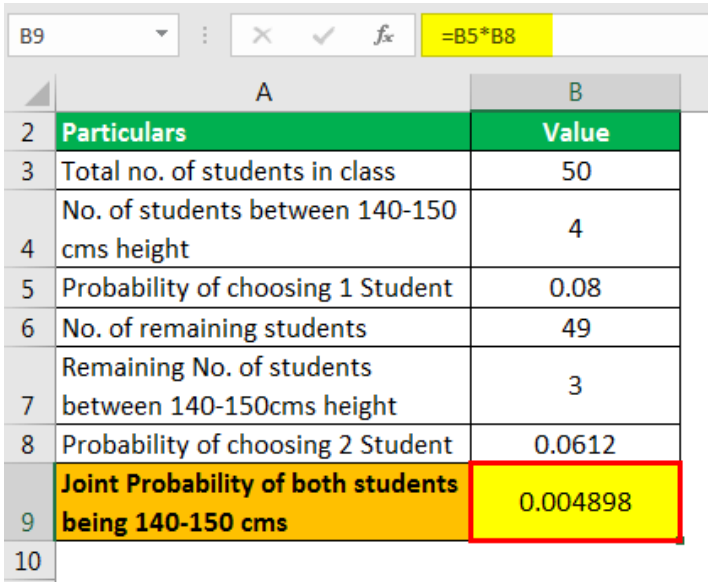
		B
2	**Particulars**	**Value**
3	Number of Red Balls	10
4	Number of Blue Balls	10
5	Probability of choosing 1 Red and 1 Blue	1
6	Possible Outcomes	4
7	Probability of choosing Red Ball	0.25
8	Probability of choosing Blue Ball	0.25
9	**Joint Probability of choosing both Red & Blue balls (JP)**	=B7*B8
10		

- =0.25*0.25



B9 =B7*B8

	A	B
2	**Particulars**	**Value**
3	Number of Red Balls	10
4	Number of Blue Balls	10
5	Probability of choosing 1 Red and 1 Blue	1
6	Possible Outcomes	4
7	Probability of choosing Red Ball	0.25
8	Probability of choosing Blue Ball	0.25
9	**Joint Probability of choosing both Red & Blue balls (JP)**	0.0625
10		



B9 =B5*B8

	A	B
2	**Particulars**	**Value**
3	Total no. of students in class	50
4	No. of students between 140-150 cms height	4
5	Probability of choosing 1 Student	0.08
6	No. of remaining students	49
7	Remaining No. of students between 140-150cms height	3
8	Probability of choosing 2 Student	0.0612
9	**Joint Probability of both students being 140-150 cms**	0.004898
10		

Example: **Find the probability that the number three will occur twice when two dice are rolled at the same time.**

Solution:

Number of possible outcomes when a die is rolled = 6

i.e. {1, 2, 3, 4, 5, 6}

Let A be the event of occurring 3 on first die and

B be the event of occurring 3 on the second die.

Both the dice have six possible outcomes, the probability of a three occurring on each die is 1/6.

P(A) =1/6

P(B)=1/6

P(A, B) = P(A) * P(B)

P(A ,B) = 1/6 x 1/6 = 1/36

Conditional Probability

Q) Explain Conditional Probability in detail.

A) The conditional probability of an **event *B*** is the probability that the event will occur given the knowledge that an **event *A* has already occurred**.

It is denoted by **P(B|A).**

For example, the conditional probability of event *A* given event *B* is written formally as:

- P (A given B)

The "*given*" is denoted using the pipe "|" operator; for example:

- P (A | B)

The conditional probability for events *A* given event *B* is calculated as follows:

- P (A given B) = P(A and B) / P(B)

The conditional probability formula for an event that is neither mutually exclusive nor independent is:

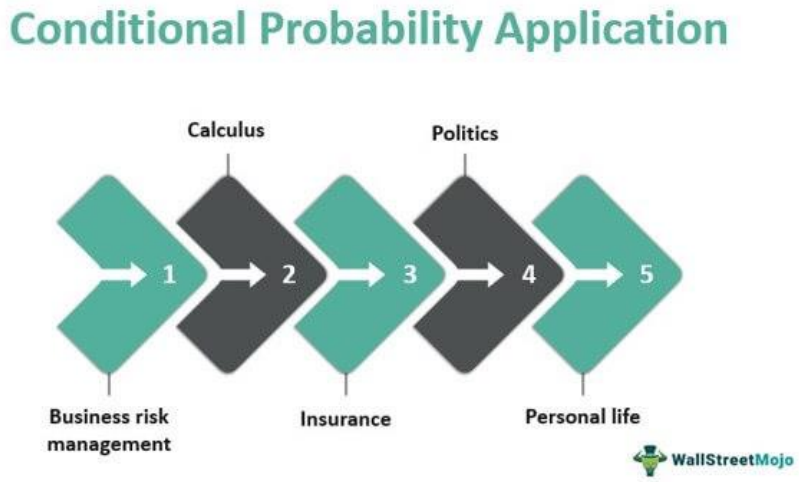
$$P(\overline{A}|B) = \frac{P(A \cap B)}{P(B)}$$

where:

- P(A|B) denotes the conditional chance, i.e., the probability of the occurrence of event A with relation to condition B.
- P(A∩B) signifies the joint probability of both events occurring. It is not what both the events cover individually but the common factor that connects both of them for the outcome.
- P(B) is the probability of B.

Conditional chance or probability can tell one about the likelihood of an outcome based on the conditions dominating a previous occurrence.

It is an essential component of probability theory. However, it is different from unconditional probability, where the chances of an event occurring do not depend on the previous outcome or condition.



Conditional Probability Example

Example #1

- B1 and B2 are the two boxes.
- The first box contains five squash balls – three blue and two green.
- The second box has seven squash balls – one blue and six green.

Karl draws a squash ball from each box randomly. How likely is Karl to choose –?

1. A blue squash ball (B)
2. A green squash ball (G)
3. Green squash ball (G) from B2

Solution

Total number of squash balls (B1+B2) = 5+7 = 12

1. The probability of a blue squash ball to be chosen = No. of blue squash balls/Total no. of squash balls

P(B) = (3+1)/12
= 4/12
= 1/3

- The probability of a green squash ball to be chosen = No. of green squash balls/Total no. of squash balls

P(G) = (2+6)/12
= 8/12
= 2/3

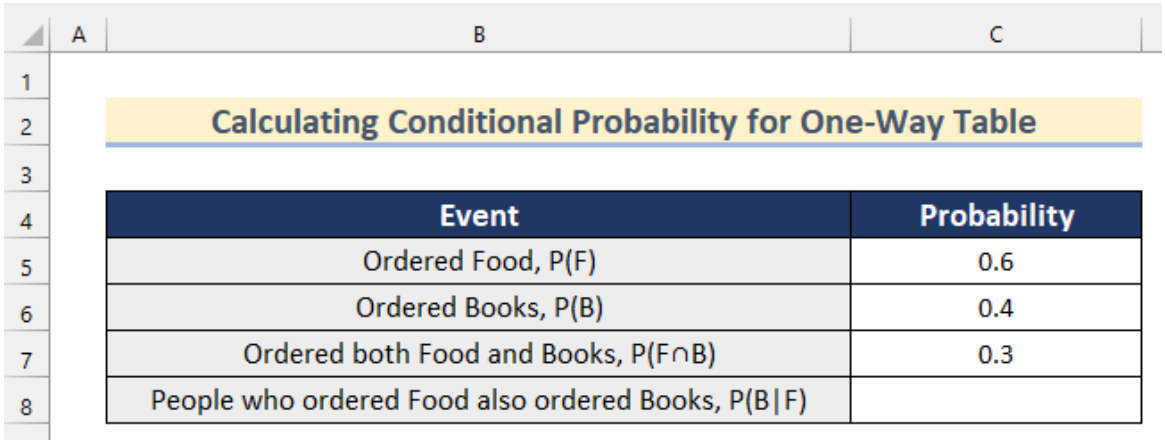
- Probability of a green squash ball to be chosen from B2 = No. of green squash balls in B2/Total no. of squash balls in B2

P(G|B2) = 6/7
The third scenario is an example of conditional chance or probability where event G needs to be figured out based on condition B2.

Calculation of Conditional Probability in Excel:

Calculate Conditional Probability for One-Way Table

In the first method, we will show you how you can **calculate the conditional probability** for a **one-way** table. Here, we have a dataset containing the **probability** of ordering **Food**, **Books** and **both** at a time. Now, using this dataset we will show you how you can **calculate** the **probability** of people who **ordered books** after **conditioning** that they have **ordered food**.



Calculating Conditional Probability for One-Way Table

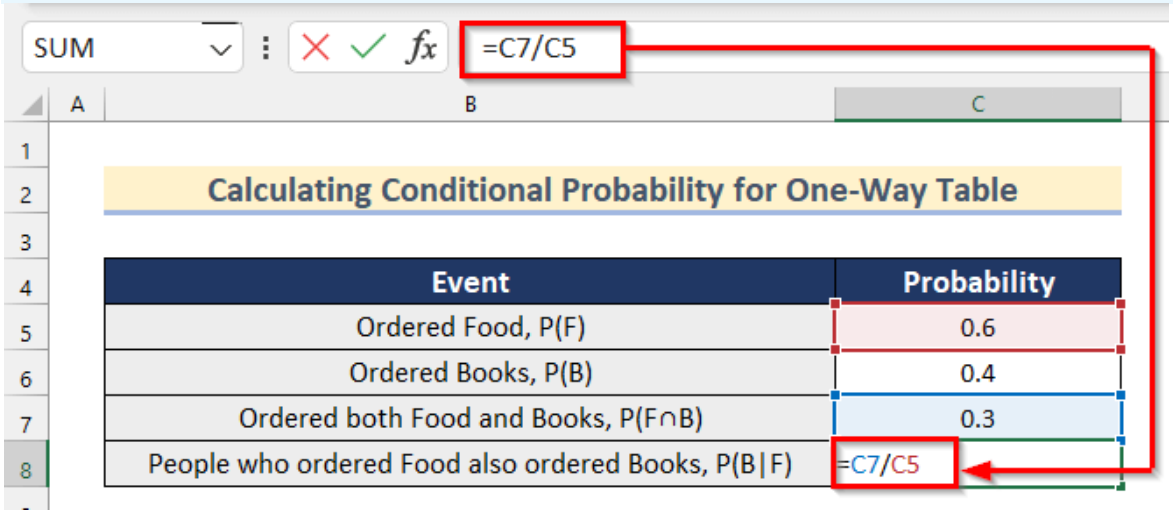
Event	Probability
Ordered Food, P(F)	0.6
Ordered Books, P(B)	0.4
Ordered both Food and Books, P(F∩B)	0.3
People who ordered Food also ordered Books, P(B|F)	

Here are the steps.

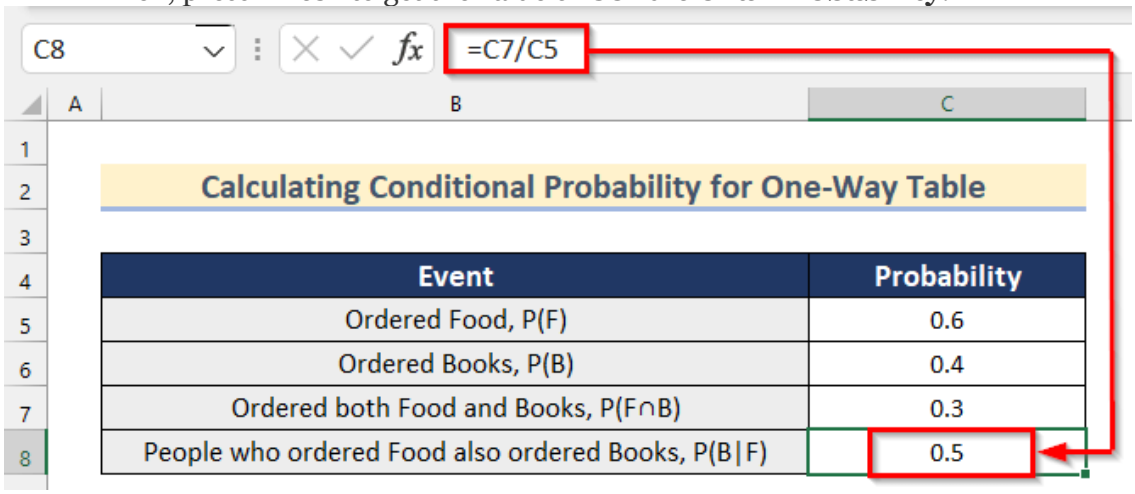
Steps:

- In the beginning, select **Cell C8** and insert the following **formula**.

=C7/C5



- Then, press **Enter** to get the value of **Conditional Probability**.



Here, in the formula, we **divided** the **probability** of **Ordering both Food** and **Books** in **Cell C7** by the **probability** of **Ordering Food** in **Cell C5** to get the value of **Conditional Probability**.

Example #2

In a book fair attended by 300 people, 130 bought fiction books (F). Seventy people chose non-fiction books (N), while 100 visitors selected books from both genres. If a randomly chosen buyer bought a fiction book, what is the probability of the same person buying a non-fiction book too?

Solution

- Probability of people choosing fiction books, i.e., P(F) = 130/300

= 0.43

- Probability of people choosing fiction and non-fiction books, i.e., P(N∩F) = 100/300

= 0.33

Probability of a random person with fiction books also choosing non-fiction, i.e.,

- Probability of a random person with fiction books also choosing non-fiction, i.e., P(N|F) = P(F∩N)/P(F)

= 0.33/0.43

= 0.767 = 0.8 (approx.)

Therefore, the probability of a random buyer choosing non-fiction books given they have already purchased a fiction book is 80%.

Calculate P using a Contingency table

https://openstax.org/books/introductory-business-statistics/pages/3-4-contingency-tables-and-probability-trees

Q) What is Contingency table?

A) A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily.

Contingency tables (also called crosstabs or two-way tables) summarize the relationship between several categorical variables. It is a special type of frequency distribution table, where two variables are shown simultaneously.

The table displays sample values in relation to two different variables that may be dependent or contingent on one another.

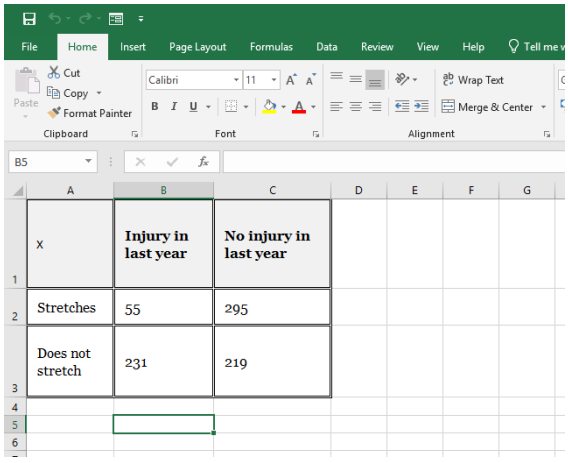
The table helps in determining conditional probabilities quite easily.

The table displays sample values in relation to two different variables that may be dependent or contingent on one another.

They are heavily used in

- **survey research,**
- **business intelligence,**
- **engineering, and**
- **scientific research.**

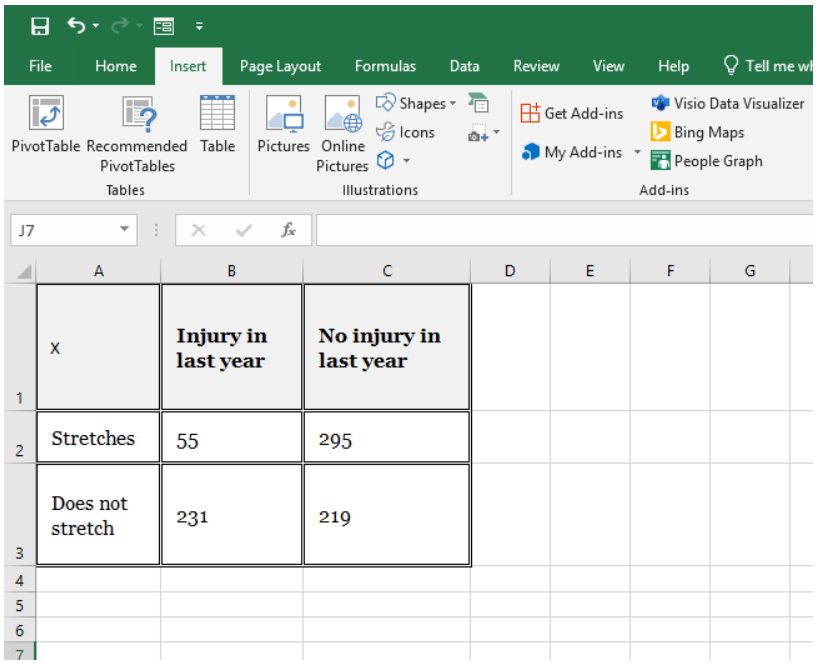
Example: CREATION OF Contingency Table in Excel



Use the following steps to create a contingency table for this dataset:

Step 1: Select the PivotTable option.

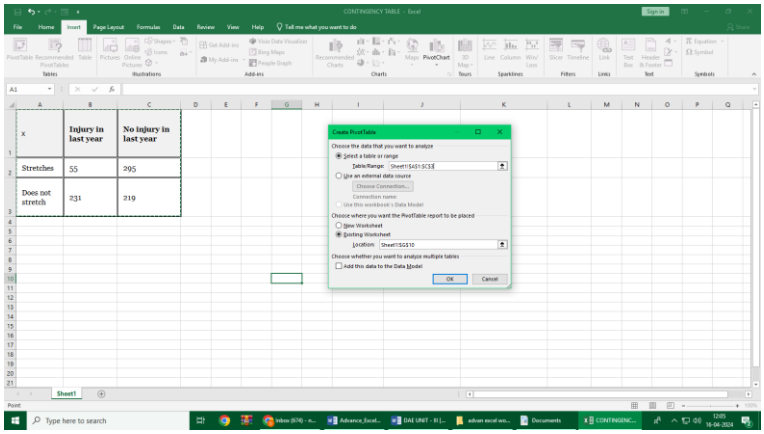
Click the **Insert** tab, then click **PivotTable**.



	A	B	C	D	E	F	G
1	X	**Injury in last year**	**No injury in last year**				
2	Stretches	55	295				
3	Does not stretch	231	219				
4							
5							
6							
7							

In the window that pops up, choose **COMPLETE TABLE** for the range of values. Then choose a location to place the pivot table in.

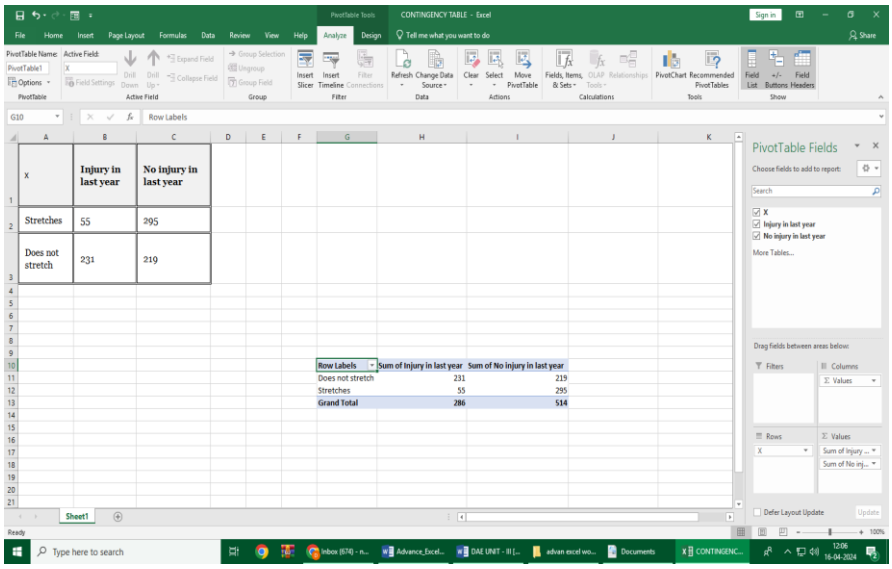
We'll choose cell **E2** within the existing worksheet:



Once you click **OK**, an empty contingency table will appear in cell E2.

Step 2: Populate the contingency table.

In the window that appears on the right, drag **X** into the box titled **Rows,** and drag **Injury in last year , No injury in last year** into the box titled **Values**:



Once you do so, the frequency values will automatically be populated in the contingency table:

Note: *If Values first appears as "Sum of Order Number" simply click the dropdown arrow and select* ***Value Field Settings****. Then choose* ***Count*** *and click* ***OK****.*

Step : 3

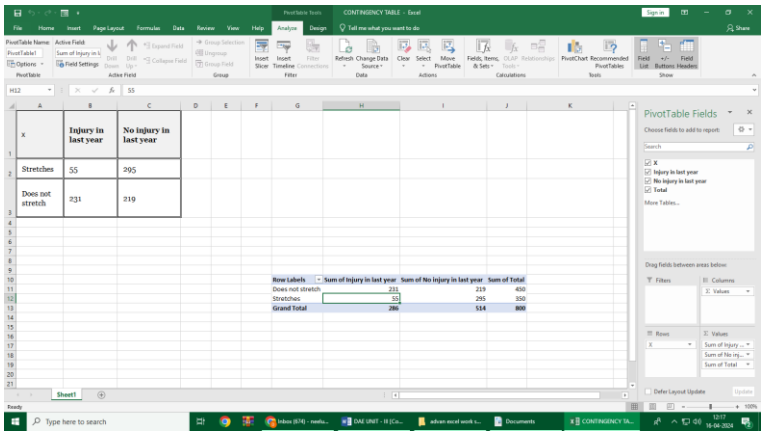
Go to Analysis tab > Fields , Items & Sets >

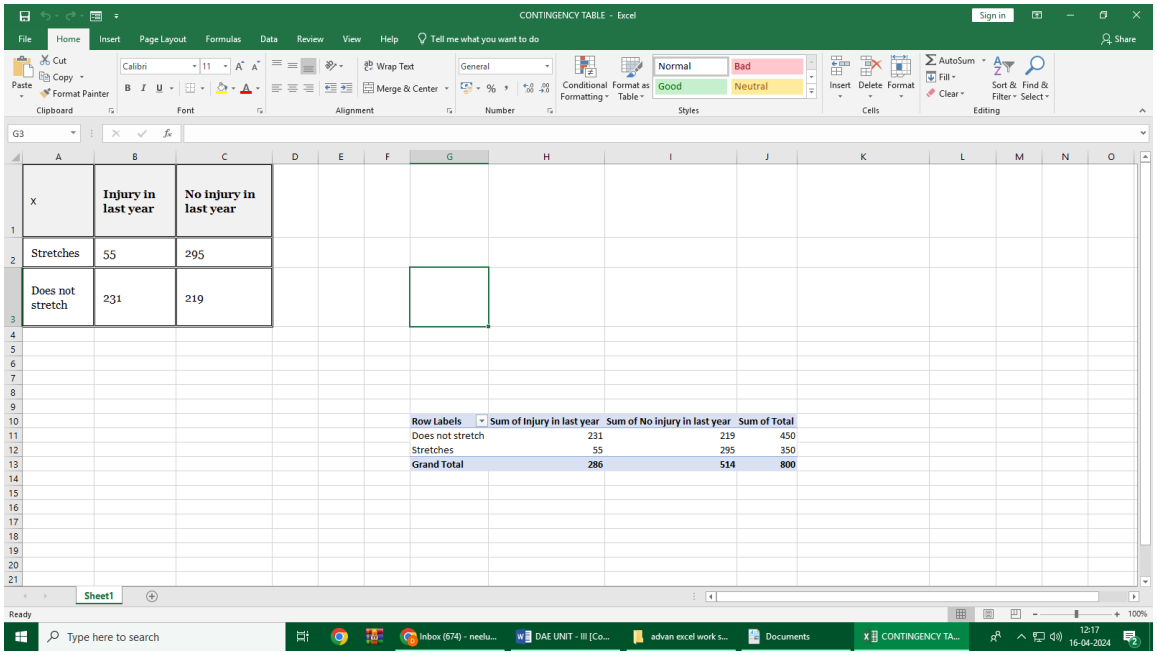
Insert calculated field Dialogue box appear

Enter user Name : Total

Enter formula : ='Injury in last year'+ 'No injury in last year'

Press ok button.





Example 1:

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

The total number of people in the sample is 755.
The row totals are 305 and 450.
The column totals are 70 and 685.
Notice that 305+450=755 and 70+685=755

Calculate the following probabilities using the table.

1. Find P (Person is a car phone user).
2. Find P (person had no violation in the last year).
3. Find P (Person had no violation in the last year AND was a car phone user).
4. Find P (Person is a car phone user OR person had no violation in the last year).
5. Find P (Person is a car phone user GIVEN person had a violation in the last year).

6. Find P (Person had no violation last year GIVEN person was not a car phone user)

Solution:

1. number of car phone users / total number in study=305 / 755
2. number that had no violation/ total number in study=685 / 755
3. 280 / 755
4. (305/755+685/755) −280 / 755=710 / 755
5. 25/ 70 (The sample space is reduced to the number of persons who had a violation
6. 405 / 450 (The sample space is reduced to the number of persons who were not car phone users.)

Example No: 2

This table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	**286**	**514**	**800**

1. What is P(athlete stretches before exercising)?
2. What is P(athlete stretches before exercising|no injury in the last year)?

Solution

1. P(athlete stretches before exercising) = 350 / 800=0.4375
2. P(athlete stretches before exercising|no injury in the last year) = 295 / 514=0.5739

Calculate P from Tree Diagrams

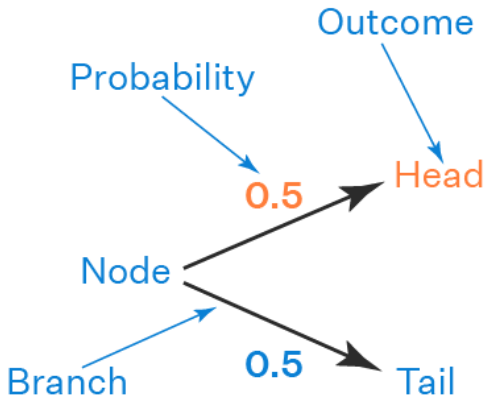
A probability tree diagram is used to represent the probability of occurrence of events without using complicated formulas. It displays all the possible outcomes of an event. The purpose of a probability tree is that it shows all the possible outcomes of an event and calculates the probability of these outcomes.
A probability tree diagram can be used to depict conditional probabilities as well as independent events

Parts of a Probability Tree Diagram

There are two main parts of a probability tree. These are the nodes and the branches. The nodes can further be classified into a parent node and a sibling node. The parent node represents a

certain event and has a probability of 1. The sibling nodes denote other additional possible events or outcomes. The branches denote the probability of occurrence of these events. Suppose a fair coin is tossed once, then the probability tree can be constructed as follows:

Probability Tree Diagram Parts

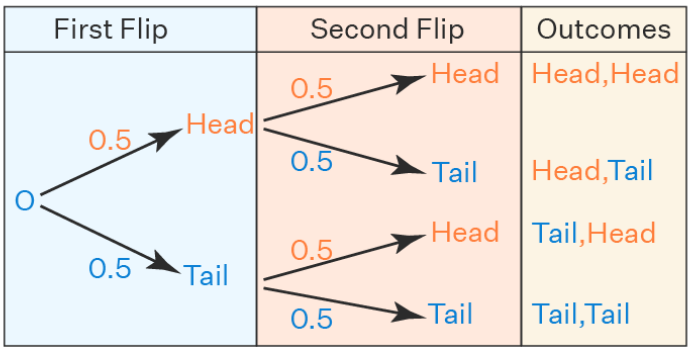


This is a simple probability tree and has two branches only. Here, the first node represents the parent event of a coin being tossed. Head and tail are the two possible outcomes forming the sibling nodes. 0.5 is written on the branch and represents the probability of occurrence of each sibling event.

Probability Tree Diagram Example

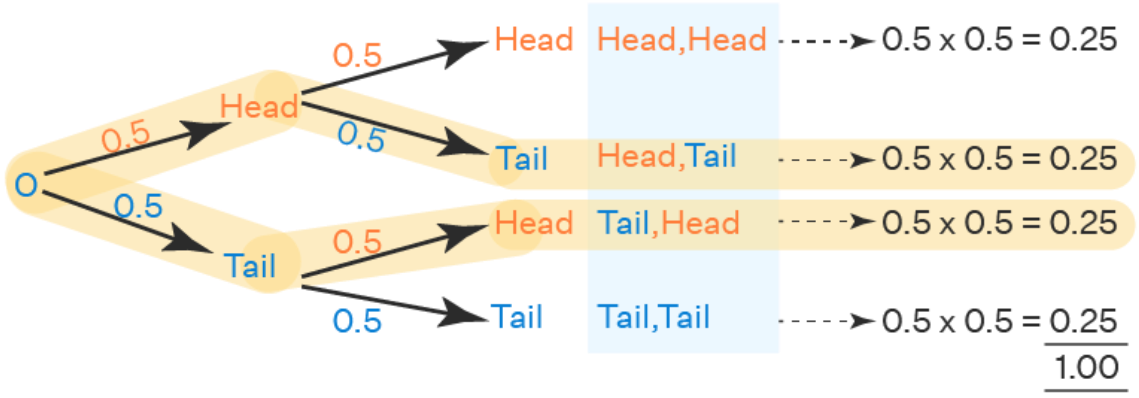
Suppose a probability tree diagram needs to be constructed for flipping a fair coin twice. This is an example of an independent event as the outcome of each coin toss is independent of the previous flip. First, the probability tree diagram of a coin being flipped once is drawn as given in the previous section. The next step is to extend it to two coin tosses as follows:

Probability Tree Diagram for a Coin Tossed Twice



The second set of probabilities represents the second coin toss. Thus, in total there are 4 sets of possible outcomes.

To calculate the probabilities of a series of events, multiply the probabilities along the branches of the probability tree diagram. The total probability can be computed by adding these probabilities and its value will always be equal to 1.



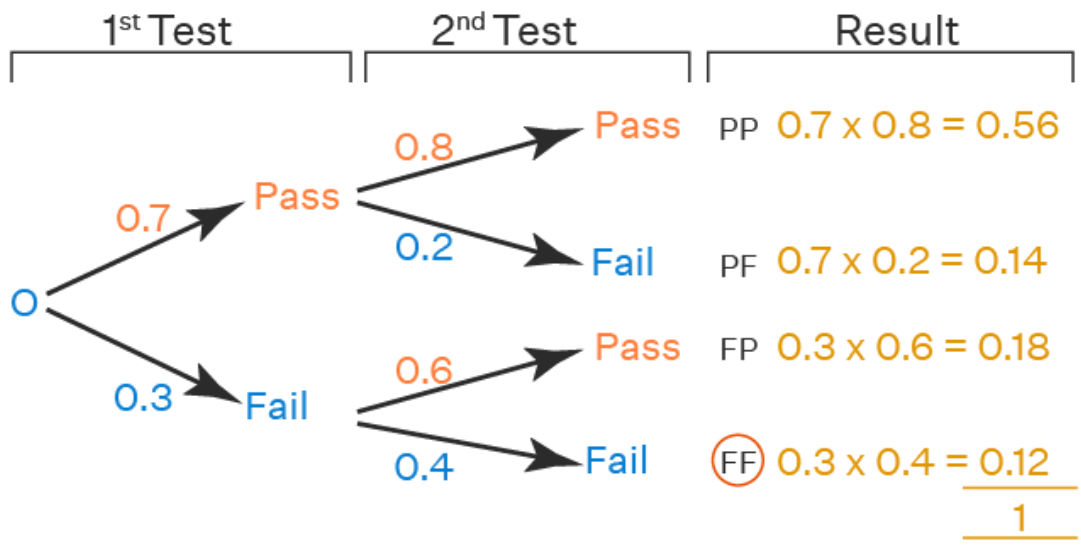
Some useful inferences can be made from the probability tree diagram as follows:

- The probability of getting the outcome (Head, Head) = 0.5 × 0.5 = 0.25. Similarly, the probability of the other outcomes can be calculated.
- 0.25 + 0.25 + 0.25 + 0.25 = 1. This implies that on adding the probabilities of each outcome, the total is equal to 1.
- By looking at the probability tree, the probability of getting exactly one head can be calculated as 0.25 + 0.25 = 0.5.

Conditional Probability Tree Diagram

A conditional probability tree diagram is very useful in depicting the outcome of dependent events. A dependent event is one whose outcome is affected by an event that has previously occurred. Suppose a student has to appear for two tests. The probability that he will pass the first test is 0.7. If he passes the first test, the probability that he passes the second test is 0.8. Furthermore, if he fails the first test, the probability that he passes the second test is 0.6. To determine the conditional probability that he fails in both tests, the probability tree diagram can be constructed using the following steps.

- Step 1: Construct the probability tree diagram for the first test. As the probability of passing the first test is 0.7 thus, the probability of failure is 1 - 0.7 = 0.3.
- Step 2: Construct branches for the second test given that the student passes the first test. It must be noted that the sum of probabilities of outcomes, 0.8 + 0.2, is 1.
- Step 3: Now, draw branches for the second test given that the student fails the first test.
- Step 4: Multiply the probabilities along each branch of the probability tree to get the conditional probability results.



From the probability tree diagram, it can be concluded that the conditional probability for a student who fails the second test given that he fails the first is 0.12.

How to Draw a Probability Tree?

To draw a probability tree diagram it is necessary to identify all the possible outcomes and the probabilities associated with them. The steps to construct a probability tree are as follows:

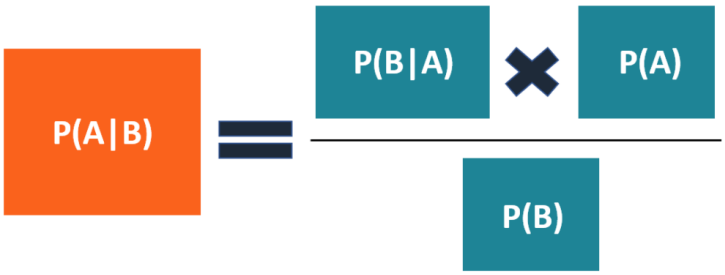
- Step 1: Identify whether the events are dependent or independent.
- Step 2: Draw branches to represent the first set of outcomes.
- Step 3: Write the probabilities associated with each outcome on the branch.
- Step 4: Draw the next set of branches taking into account whether the subsequent events are dependent or independent. Also, write the associated probabilities.
- Step 5: Repeat step 5 for as many branches as required.
- Step 6: To calculate the overall probability of the probability tree, the probabilities along the branches have to be multiplied. The sum of these probabilities should always be equal to 1.

Q) What is the Bayes' Theorem?

A) In statistics and probability theory, the Bayes' theorem (also known as the Bayes' rule) is a mathematical formula used to determine the conditional probability of events.

Essentially, the Bayes' theorem describes the probability of an event based on prior knowledge of the conditions that might be relevant to the event.

The theorem is named after English statistician, Thomas Bayes, who discovered the formula in 1763. It is considered the foundation of the special statistical inference approach called the Bayes' inference.



Besides statistics, the Bayes' theorem is also used in various disciplines, with medicine and pharmacology as the most notable examples. In addition, the theorem is commonly employed in different fields of finance. Some of the applications include but are not limited to, modeling the risk of lending money to borrowers or forecasting the probability of the success of an investment.

Formula for Bayes' Theorem

The Bayes' theorem is expressed in the following formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:

- P(A|B) – the probability of event A occurring, given event B has occurred
- P(B|A) – the probability of event B occurring, given event A has occurred
- P(A) – the probability of event A
- P(B) – the probability of event B

Note that events A and B are independent events (i.e., the probability of the outcome of event A does not depend on the probability of the outcome of event B).

Important Notes on Bayes Theorem

- Bayes theorem is used to determine conditional probability.
- When two events A and B are independent, P(A|B) = P(A) and P(B|A) = P(B)
- Conditional probability can be calculated using the Bayes theorem for continuous random variables.

Example of Bayes' Theorem

Imagine you are a financial analyst at an investment bank. According to your research of publicly-traded companies, 60% of the companies that increased their share price by more than 5% in the last three years replaced their CEOs during the period.

At the same time, only 35% of the companies that did not increase their share price by more than 5% in the same period replaced their CEOs. Knowing that the probability that the stock prices grow

by more than 5% is 4%, find the probability that the shares of a company that fires its CEO will increase by more than 5%.

Before finding the probabilities, you must first define the notation of the probabilities.

- P(A) – the probability that the stock price increases by 5%
- P(B) – the probability that the CEO is replaced
- P(A|B) – the probability of the stock price increases by 5% given that the CEO has been replaced
- P(B|A) – the probability of the CEO replacement given the stock price has increased by 5%.

Using the Bayes' theorem, we can find the required probability:

$$P(A|B) = \frac{0.60 \times 0.04}{0.60 \times 0.04 + 0.35 \times (1 - 0.04)} = 0.067 \text{ or } 6.67\%$$

Thus, the probability that the shares of a company that replaces its CEO will grow by more than 5% is 6.67%.

For example, suppose the probability of the weather being cloudy is 40%. Also suppose the probability of rain on a given day is 20% and that the probability of clouds on a rainy day is 85%.

If it's cloudy outside on a given day, what is the probability that it will rain that day?

Solution:

- P(cloudy) = 0.40
- P(rain) = 0.20
- P(cloudy | rain) = 0.85

Thus, we can calculate:

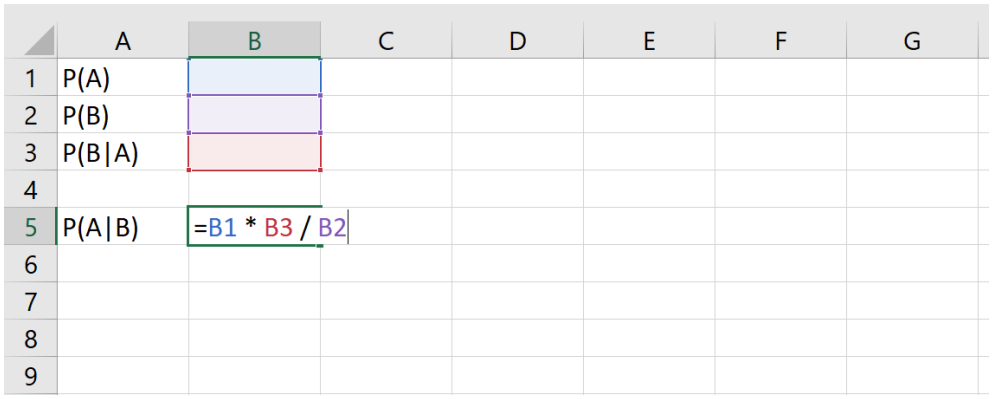
- P(rain | cloudy) = P(rain) * P(cloudy | rain) / P(cloudy)
- P(rain | cloudy) = 0.20 * 0.85 / 0.40
- P(rain | cloudy) = 0.425

If it's cloudy outside on a given day, the probability that it will rain that day is **0.425** or **42.5%**.

The following example shows how to solve this exact problem using Bayes' Theorem in Excel.

Example: Bayes' Theorem in Excel

The following formula shows how to apply Bayes' Theorem in Excel:



	A	B	C	D	E	F	G
1	P(A)						
2	P(B)						
3	P(B|A)						
4							
5	P(A|B)	=B1 * B3 / B2					
6							
7							
8							
9							

For example, if we know the following probabilities:

- P(cloudy) = 0.40
- P(rain) = 0.20
- P(cloudy | rain) = 0.85

Then we can simply plug these into the cells in Excel:

	A	B	C	D	E	F
1	P(A)	0.2		P(rain) = 0.20		
2	P(B)	0.4		P(cloudy) = 0.40		
3	P(B|A)	0.85		P(cloudy|rain) = 0.85		
4						
5	P(A|B)	**0.425**		P(rain|cloudy)		
6						
7						
8						
9						
10						

This tells us that if it's cloudy outside on a given day, the probability that it will rain that day is **0.425** or **42.5%**.

Example2

- Dangerous fires are rare (1%)
- But smoke is fairly common (10%) due to barbecues,
- And 90% of dangerous fires make smoke

What is the probability of dangerous Fire when there is Smoke?

Calculation

P(Fire|Smoke) =P(Fire) P(Smoke|Fire)/P(Smoke)
= 1% x 90%/10%
= 9%

Difference Between Conditional Probability and Bayes Theorem

Conditional Probability	Bayes Theorem
Conditional Probability is the probability of an event A that is based on the occurrence of another event B.	Bayes Theorem is derived using the definition of conditional probability. The Bayes theorem formula includes two conditional probabilities.
$P(A\|B) = \frac{P(A \cap B)}{P(B)}$	$P(A\|B) = \frac{P(B\|A)P(A)}{P(B)}$

Q) Calculation of Mean, variance, standard deviation in terms of probabilities in excel

Ref : https://www.youtube.com/watch?v=Lvk2wmUy21E

A probability distribution tells us the probability that a random variable takes on certain values.

To find the **mean** (sometimes called the "expected value") of any probability distribution, we can use the following formula:

Method 1 :Mean (Or "Expected Value") of a Probability Distribution:

$$\mu = \Sigma(x * P(x))$$

where:

- **x: Data value**
- **P(x): Probability of value**

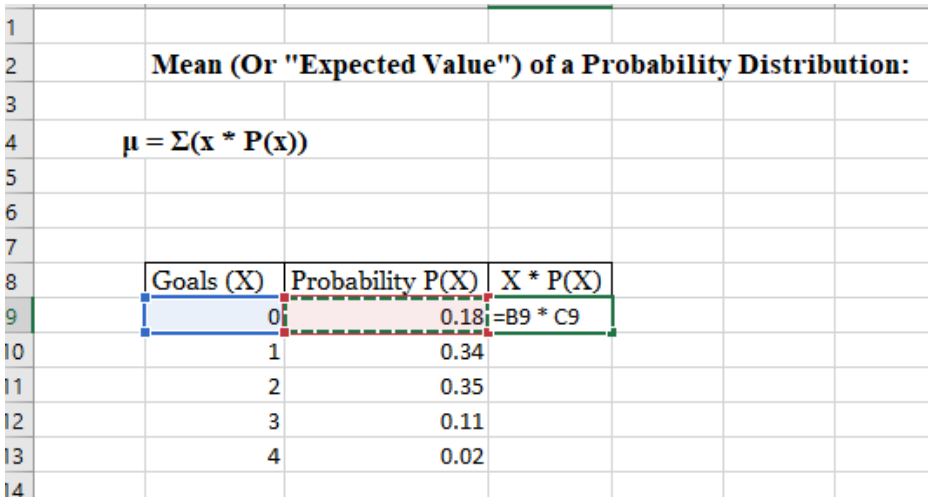
Method 2 : Ms- Excel formula – average ( )

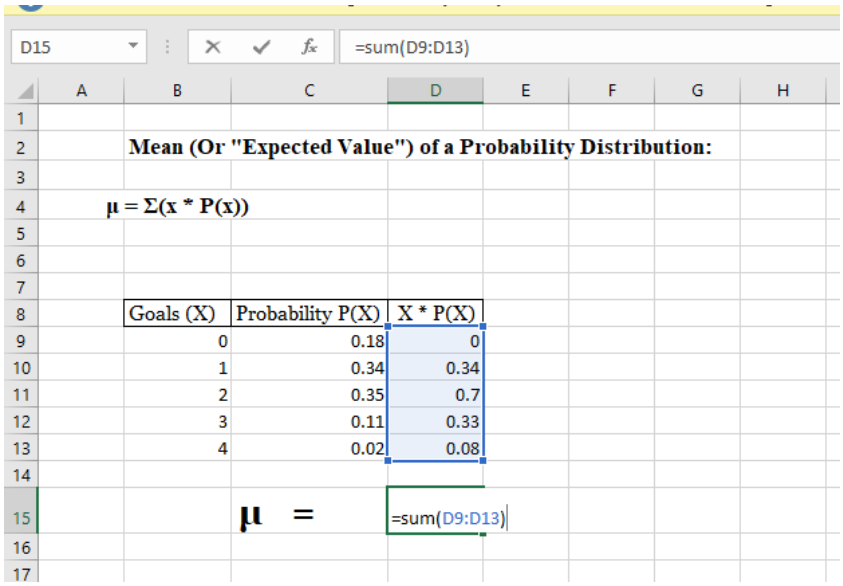
Method 1:
Example :

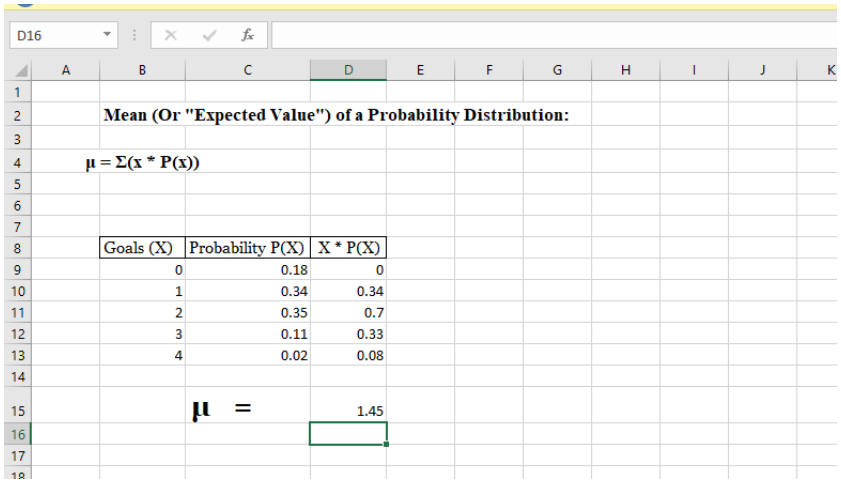
Goals (X)	Probability P(X)
0	0.18
1	0.34
2	0.35
3	0.11
4	0.02

Steps :

1) Enter data
2) Type X * P(X) in adjacent cell to P(X)
3) Put =B9 * C9, press enter
4) Press auto fill for next value
5) Select any cell down the data , put =Sum( ) and select the range i.e D9-D13, enter

- 

- 



	A	B	C	D	E	F	G	H	I	J	K
1											
2		Mean (Or "Expected Value") of a Probability Distribution:									
3											
4		μ = Σ(x * P(x))									
5											
6											
7											
8		Goals (X)	Probability P(X)	X * P(X)							
9		0	0.18	0							
10		1	0.34	0.34							
11		2	0.35	0.7							
12		3	0.11	0.33							
13		4	0.02	0.08							
14											
15			μ =	1.45							
16											
17											

Method 2 :

1) **Enter data**
2) **Select one cell type mean and in adjacent cell type = average ()**
3) **Select the range of which it**

Calculation of variance & standard deviation in terms of probabilities in excel

Standard deviation of a probability distribution measures the scattering of the probability distribution with respect to its mean. It is a measure obtained by taking the square root of the variance.

METHOD 1 : CONVENTIONAL FORMULA

METHOD 2 : VAR(), STDEV() refer unit - II

Method 1 : conventional formula :

The following sequential steps help in easily computing the standard deviation of a probability distribution.

Probability Distribution Formula

Mean ($\bar{x}$) = $\sum [x_i * P(x_i)]$

Standard Deviation (σ) = $\sqrt{\sum (x_i - \bar{x})^2 * P(x_i)}$

Where,

- x_i = Value of the random variable in the i^{th} observation
- $P(x_i)$ = Probability of the i^{th} value

Step 1: First, determine the values of the random variable or event through several observations, denoted by x_1, x_2, ..., x_n, or x_i.

Step 2: Next, compute the probability of occurrence of each value of the random variable, and they are denoted by $P(x_1)$, $P(x_2)$, ..., $P(x_n)$, or $P(x_i)$.

$$P(x_i) = \text{No. of Events with } i^{th} \text{ Value / Total No. of Events}$$

Step 3: The formula for the mean can be derived by adding up the products of the value of the random variable (step 1) and its probability (step 2), as shown below:

$$\bar{x} = \sum[x_i * P(x_i)]$$

Step 4: Compute each value's deviation (step 1) of the random variable from the probability distribution's mean (step 3).

$$D_i = x_i - \bar{x}$$

Step 5: Next, The formula for standard deviation can be derived by adding up the products Of the squares of the deviation of each value (step 4) and its corresponding probability (step 2) and then computing the square root of the result, as shown below.

$$\sigma = \sqrt{[\sum(x_i - \bar{x})^2 * P(x_i)]}$$

Example #1

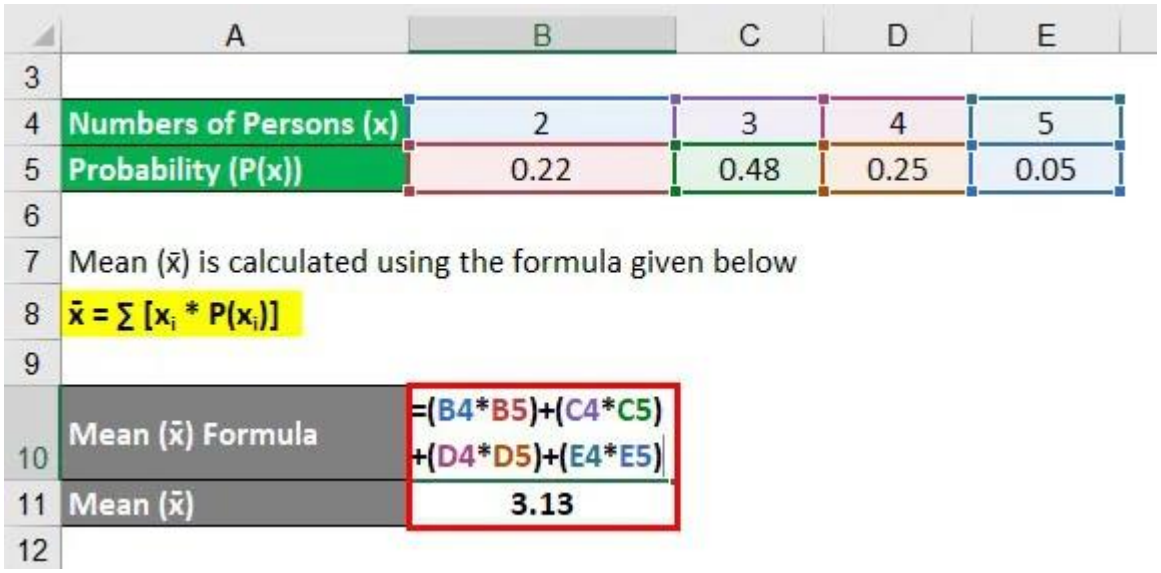
Let's take the example of a survey conducted in a certain area to find out the expected number of persons in a family. The following data is available. Calculate the mean and standard deviation of the probability distribution.

	A	B	C	D	E
3					
4	Numbers of Persons (x)	2	3	4	5
5	Probability (P(x))	0.22	0.48	0.25	0.05
6					

Solution:

Mean ($\bar{x}$) is calculated using the formula given below:

$$\bar{x} = \sum [x_i * P(x_i)]$$

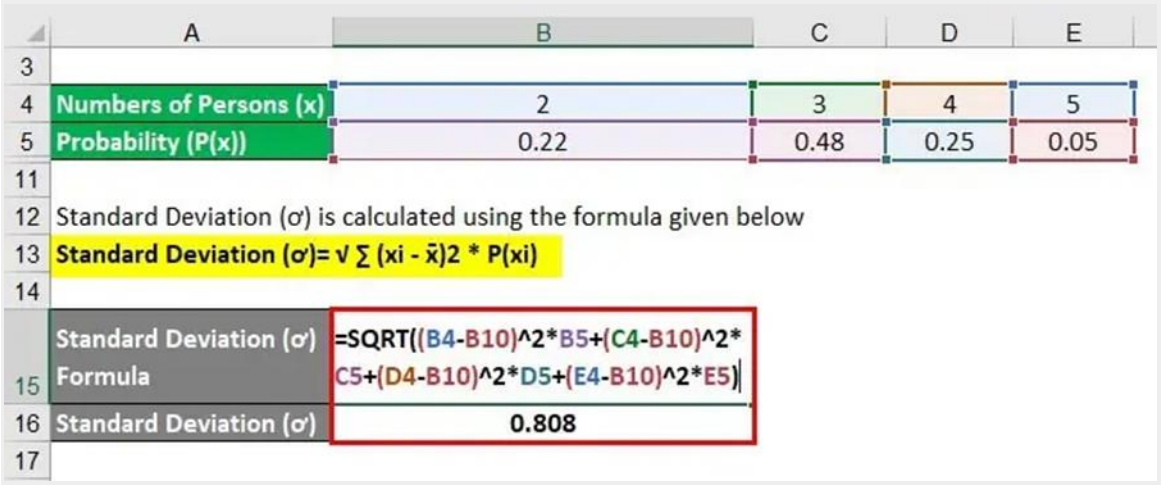


	A	B	C	D	E
3					
4	Numbers of Persons (x)	2	3	4	5
5	Probability (P(x))	0.22	0.48	0.25	0.05
6					
7	Mean ($\bar{x}$) is calculated using the formula given below				
8	$\bar{x} = \sum [x_i * P(x_i)]$				
9					
10	Mean ($\bar{x}$) Formula	=(B4*B5)+(C4*C5)+(D4*D5)+(E4*E5)			
11	Mean ($\bar{x}$)	3.13			
12					

- Mean ($\bar{x}$) = 2 * 0.22 + 3 * 0.48 + 4 * 0.25 + 5 * 0.05
- Mean ($\bar{x}$) = **3.13**

The Standard Deviation (σ) is calculated using the formula given below

Standard Deviation (σ)= $\sqrt{\sum (x_i - \bar{x})^2 * P(x_i)}$

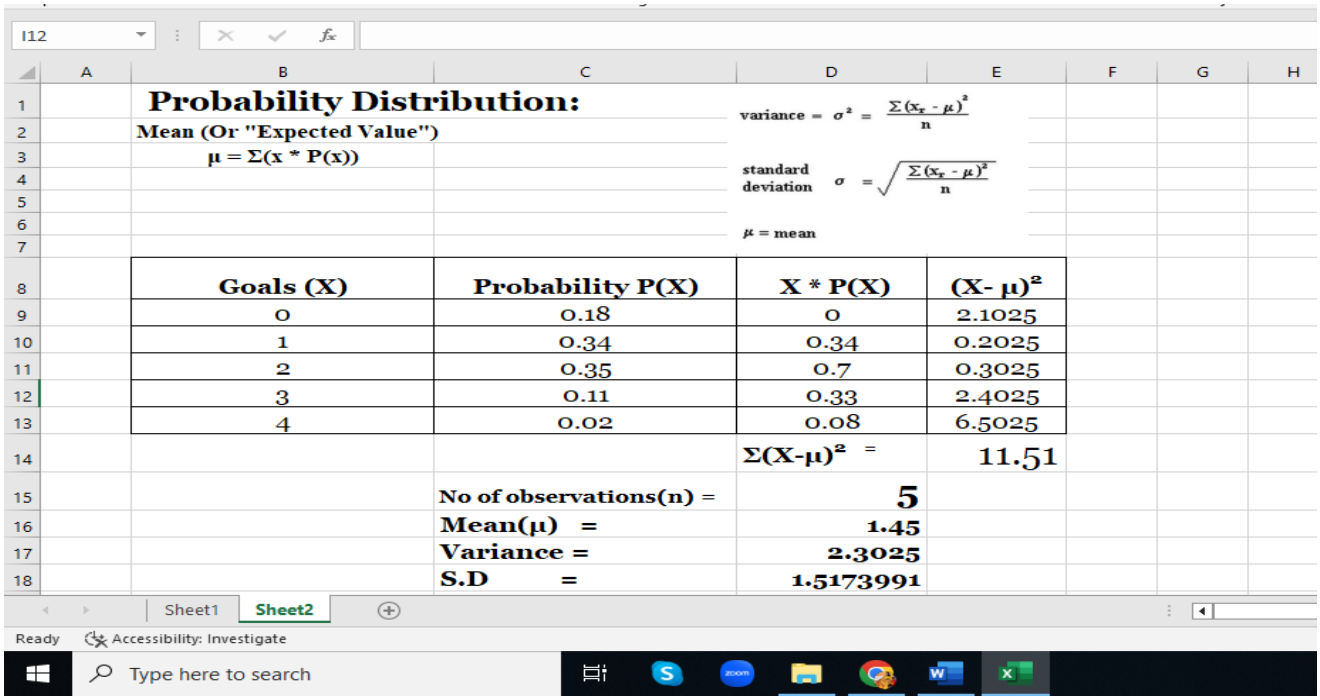


	A	B	C	D	E
3					
4	Numbers of Persons (x)	2	3	4	5
5	Probability (P(x))	0.22	0.48	0.25	0.05
11					
12	Standard Deviation (σ) is calculated using the formula given below				
13	Standard Deviation (σ)= √ ∑ (xi - x̄)2 * P(xi)				
14					
15	Standard Deviation (σ) Formula	=SQRT((B4-B10)^2*B5+(C4-B10)^2*C5+(D4-B10)^2*D5+(E4-B10)^2*E5)			
16	Standard Deviation (σ)	0.808			
17					

- **Standard Deviation (σ)=$\sqrt{[(2 - 3.13)^2 * 0.22 + (3 - 3.13)^2 * 0.48 + (4 - 3.13)^2 * 0.25 + (5 - 3.13)^2 * 0.05]}$**
- = **0.808**

EXAMPLE 1 :

Goals (X)	Probability P(X)
0	0.18
1	0.34
2	0.35
3	0.11
4	0.02



Probability Distribution:

Mean (Or "Expected Value")

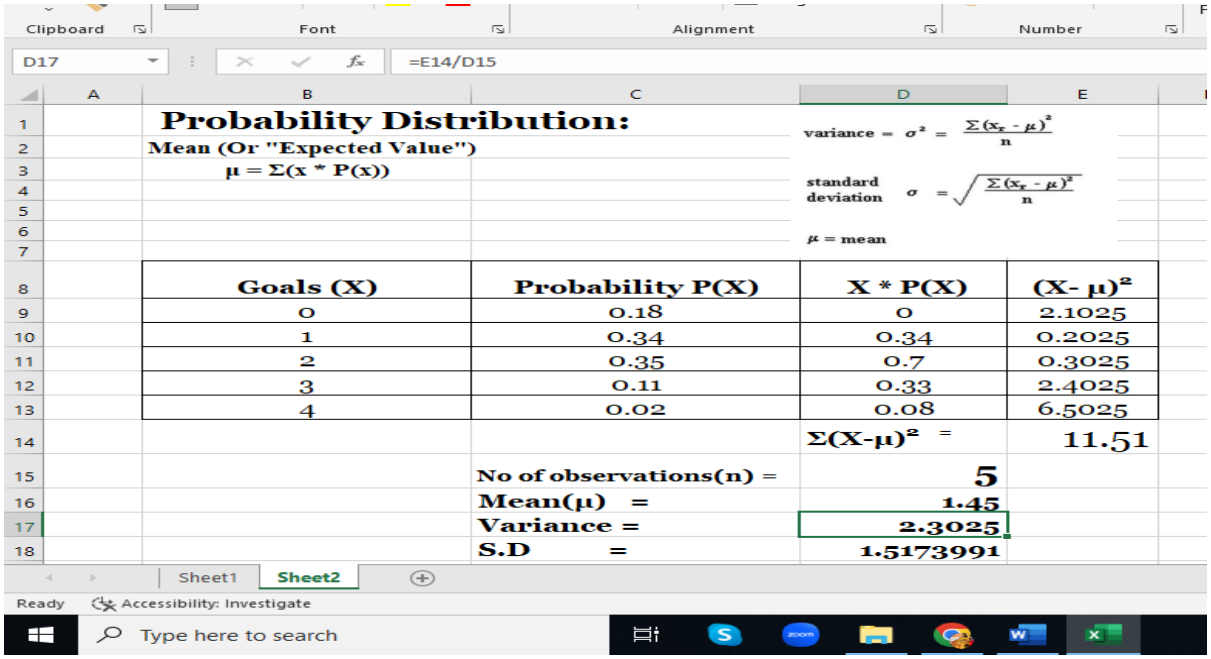
$\mu = \Sigma(x * P(x))$

variance $= \sigma^2 = \frac{\Sigma(x_x - \mu)^2}{n}$

standard deviation $\sigma = \sqrt{\frac{\Sigma(x_x - \mu)^2}{n}}$

μ = mean

Goals (X)	Probability P(X)	X * P(X)	$(X-\mu)^2$
0	0.18	0	2.1025
1	0.34	0.34	0.2025
2	0.35	0.7	0.3025
3	0.11	0.33	2.4025
4	0.02	0.08	6.5025
		$\Sigma(X-\mu)^2$ =	11.51
	No of observations(n) =	5	
	Mean(μ) =	1.45	
	Variance =	2.3025	
	S.D =	1.5173991	



Clipboard | Font | Alignment | Number

D17 | =E14/D15

Probability Distribution:

Mean (Or "Expected Value")

$\mu = \Sigma(x * P(x))$

variance = $\sigma^2 = \frac{\Sigma(x_r - \mu)^2}{n}$

standard deviation $\sigma = \sqrt{\frac{\Sigma(x_r - \mu)^2}{n}}$

μ = mean

Goals (X)	Probability P(X)	X * P(X)	$(X-\mu)^2$
0	0.18	0	2.1025
1	0.34	0.34	0.2025
2	0.35	0.7	0.3025
3	0.11	0.33	2.4025
4	0.02	0.08	6.5025
		$\Sigma(X-\mu)^2$ =	11.51
	No of observations(n) =	5	
	Mean(μ) =	1.45	
	Variance =	2.3025	
	S.D =	1.5173991	

Sheet1 | Sheet2

Ready | Accessibility: Investigate

Type here to search

UNIT – IV

Distributions: Analyze distributions - Discrete distributions - Binomial distributions – Poisson distributions - Continuous Distributions: Identify continuous distributions - Calculate continuous distributions - Identify cumulative distributions - Identify normal distributions - Calculate normal distributions - Compare quartiles and normal distributions - Identify skew

Q) What is Distribution? Analyse types of Distributions

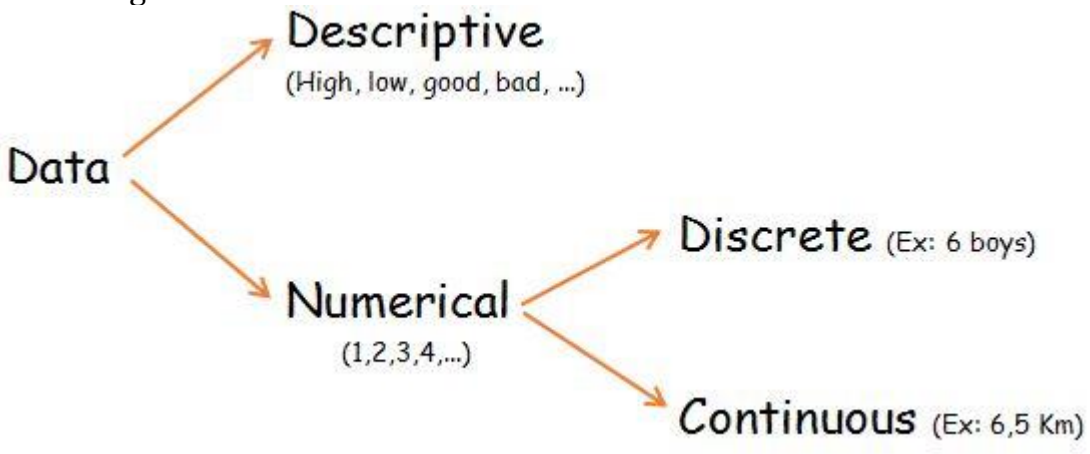
A) The distribution of a statistical dataset is the spread of the data which shows all possible values or intervals of the data and how they occur.

A distribution is simply a collection of data or scores on a variable. Usually, these scores are arranged in order from ascending to descending and then they can be presented graphically.

Difference between Frequency and Probability Distribution

S.No	Frequency Distribution	Probability Distribution
1	It records how often an event occurs. It is based on actual observations	It records the likelihood that an event is to occur. It is based on theoretical assumption of what should happen

The mean (average), standard deviation, skewness, and kurtosis of the distribution are among these factors.



Discrete data only includes values that can only be counted in integers or whole numbers. So they cannot be broken down into decimal or fraction.

Discrete Data can only take certain values

Example

The number of boys in a class is 6. (We can't have half a boy)

Definition of Continuous Data
Continuous data is the data that can be measured on a scale. It can take any numeric value, within a finite or infinite range of possible value.
he continuous data can be broken down into decimal and fractions ,so it can be subdivided into smaller parts according to the measurement precision.

Continuous Data can take any value (within a range)

Example

:
A distance: could be any value (within the range of distance such as, Km, m, Cm, …), not just certain distance.

Q) What is Probability Distribution?

A)
• It is a table of possible outcomes of an experiment and the probability of each outcome.
• It describes the chance of the future event
• The probability of each outcome is between 0 and 1 •
The sum of all mutually exclusive possible outcomes is

The Role of Probability Distribution in Business Management

1. Scenario Analysis

Probability distributions can be used to create scenario analyses. A scenario analysis uses probability distributions to create several, theoretically distinct possibilities for the outcome of a particular course of action or future event.

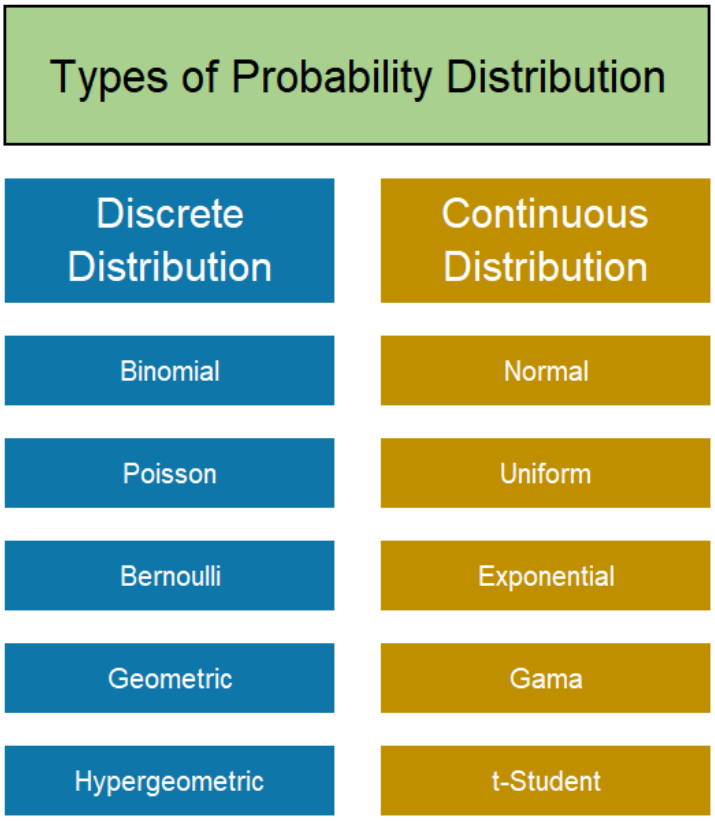
2. **Sales Forecasting**

One practical use for probability distributions and scenario analysis in business is to predict future levels of sales. It is essentially impossible to predict the precise value of a future sales level; however, businesses still need to be able to plan for future events.

3. Risk Evaluation

In addition to predicting future sales levels, probability distribution can be a useful tool for evaluating risk.

Types of Probability Distribution



The probability distribution is divided into two parts:

1. **Discrete Probability Distributions**
   - **Poisson distributions**
   - **Binomial Distributions**

2. **Continuous Probability Distributions**
   - **Normal distribution**

Discrete Probability Distribution

1. Discrete Probability Distributions

These distributions model the probabilities of random variables that can have discrete values as outcomes. For example, the possible values for the random variable X that represents the number of heads that can occur when a coin is tossed twice are the set {0, 1, 2} and not any value from 0 to 2 like 0.1 or 1.6.

Examples: poisson , Bernoulli, Binomial, Negative Binomial, Hypergeometric,

Types of Discrete probability Distributions:

Poisson Distributions

Q) Explain POISSON Distribution.

A) Poisson distribution is used to estimate how many times an event is likely to occur within the given period of time. λ is the Poisson rate parameter that indicates the expected value of the average number of events in the fixed time interval.
Poisson distribution has wide use in the fields of business as well as in biology.

Poisson distribution formula

The probability mass function of the Poisson distribution is:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Where:

- X is a random variable following a Poisson distribution
- k is the number of times an event occurs
- $P(X = k$) is the probability that an event will occur *k* times
- e is Euler's constant (approximately 2.718)
- λ is the average number of times an event occurs
- ! is the factorial function

Applications of Poisson Distribution

- **Telecommunications/management**: The number of phone calls received at a call center within a minute.
- **Staffing/scheduling**: The optimal number of agents to handle customer service calls during peak hours.
- **Transportation/operations**: The number of cars randomly arriving at a car wash in a 20-minute period.
- **Sales/marketing**: The number of customers per minute entering a bookstore.
- **Sports/gaming**: The number of goals scored in a World Cup game.
- **Biology**: The number of mutations on a strand of DNA after a limited exposure to radiation.

Uses of Poisson distribution

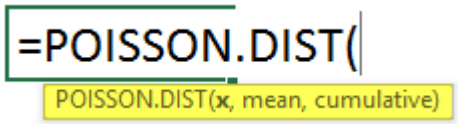
The Poisson distribution has various uses.

- Call centers use the Poisson distribution to predict how many calls they will receive per hour, allowing them to determine how many call centre representatives to keep on staff.

- Website hosting companies use the Poisson distribution to predict how many visitors a website will receive per hour.
- Restaurants use the Poisson distribution to predict the number of customers who will visit them each day.
- The Poisson distribution is used by banks to model the number of expected customer bankruptcies per month.
- The Poisson distribution is used by technology companies to model the number of expected network failures per week.

Calculation of Poisson Distribution using Microsoft Excel :

The syntax or formula for the Poisson distribution function in Microsoft Excel is:



The **POISSON.DIST( )** function syntax or formula has the below-mentioned argument:

- **x:** it is the total number of events whose probability of occurrences will be calculated.

Note: It should not be a negative value; it must be ≥ 0).
This value should be an integer; If a decimal is supplied, it will be truncated to an integer by Excel.

- **Mean**: The expected number of events to occur (Note: it must be ≥ 0).
- **Cumulative**: A logical argument that specifies the type of distribution to be calculated.

Here the type of distribution may be, or It can be either of the below mentioned:

- **TRUE or 1** – Use the cumulative distribution function or

It will return the cumulative probability of the event x or less happening.

- **FALSE or 0** – Use the probability mass or density function.

i.e., Excel will return the probability of only x number of events happening.

Examples of Poisson Distribution

- Number of Network Failures per Week
- Number of Bankruptcies Filed per Month
- Number of Website Visitors per Hour

- Number of Arrivals at a Restaurant .
- Number of Calls per Hour at a Call Center
- Number of Books Sold per Week
- Average Number of Storms in a City
- Number of Emergency Calls Received by a Hospital Every Minute

Example 1: Probability of a specific number of events

Suppose you have a call center that receives an average of 10 calls per hour. You want to calculate the probability of receiving exactly 7 calls in an hour.

You can use the POISSON.DIST function as follows:

=**POISSON.DIST**(7, 10, FALSE)

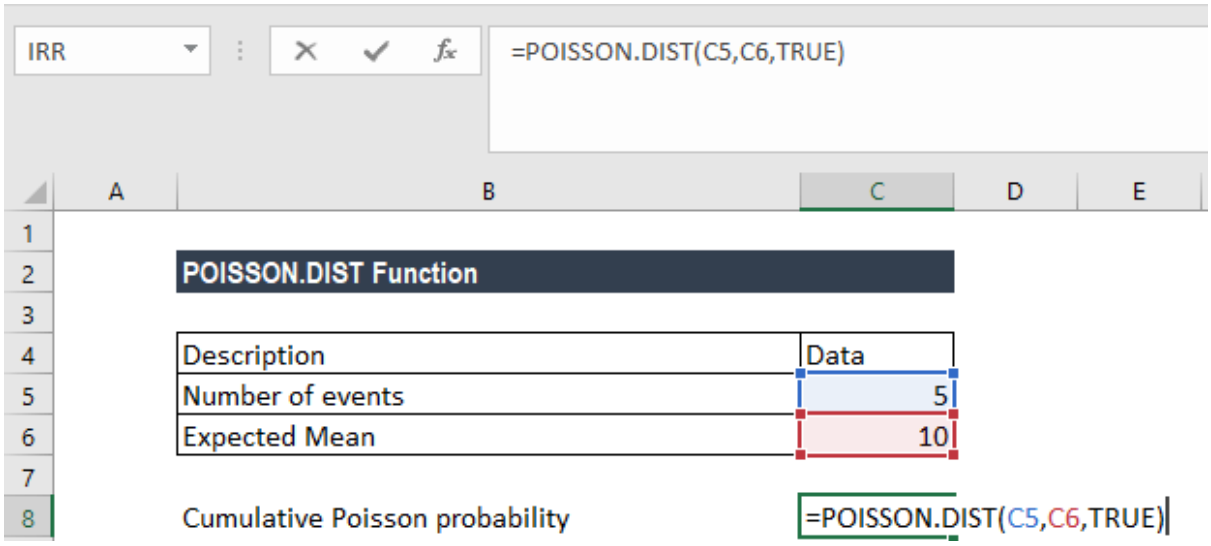
This formula will return the probability of receiving exactly 7 calls in an hour, given the average rate of 10 calls per hour.

Example 2:

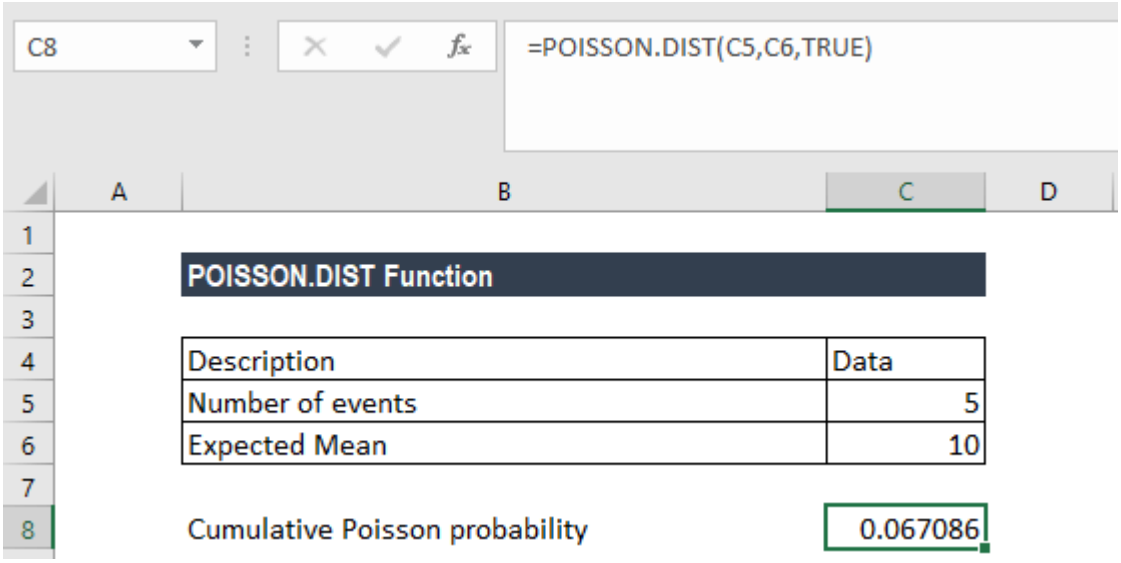
Suppose we are given the following data:

- Number of events: **5**
- Expected mean: **10**

To find out the Cumulative Poisson probability, we will use the following formula:



We get the result below:



For more examples https://www.exceldemy.com/poisson-probability-distribution-excel/

Binomial Distributions

Q) What is Binomial Distributions? How we Analyse binomial Distribution in Excel.

or

Calculate Binomial Distribution in Excel

A) Binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has **two possible outcomes** (the prefix "bi" means two, or twice).

Example are options such as:

- PASS / FAIL
- GO / NO-GO
- SUCCESS / FAILURE
- IN / OUT
- YES / NO
- HOT / COLD
- MALE / FEMALE
- WORKS / DOESN'T WORK
- DEFECTIVE / NOT DEFECTIVE
- RIGHT-HANDED / LEFT-HANDED
- HIGH / LOW
- CORRECT / INCORRECT
- HEADS / TAILS
- A coin toss has only two possible outcomes: heads or tails and

- taking a test could have two possible outcomes: pass or fail.

Applications Of Binomial Distributions :

- Testing a Drug.
- Participating in a Lucky Draw.
- Estimate the Number of Fraudulent Transactions.
- Number of Spam Emails Received.
- Number of Returns.
- Participating in an Election.
- Supporting a Particular Sports Team

Binomial Distribution Formula

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

where
n = the number of trials (or the number being sampled)
x = the number of successes desired
p = probability of getting a success in one trial
$q = 1 - p$ = the probability of getting a failure in one trial

Mean and Variance of Binomial Distribution

Mean (μ): The mean represents the average number of successes in a binomial distribution. It is calculated by multiplying the number of trials (n) by the probability of success (p).

$$\mu = np$$

Variance (σ^2): The variance measures how much the number of successes varies from the mean in a binomial distribution.

$$\sigma^2 = npq$$

Shape of Binomial Distribution

Binomial Distribution may be symmetrical or skewed.

Binomial Distribution calculation using formulas and Sample example.

Calculation of Binomial Distribution in Excel

The **BINOM.DIST** Function[1] is categorized under Excel Statistical functions.

It calculates the binomial distribution probability for the number of successes from a specified number of trials. This binomial distribution Excel guide will show you how to use the function, step by step.

The binomial distribution is a statistical measure that is frequently used to indicate the probability of a specific number of successes occurring from a specific number of independent trials.

The two forms used are:

- The Probability Mass Function – Calculates the probability of there being exactly x successes from n independent trials
- The Cumulative Distribution Function – Calculates the probability of there being at most x successes from n independent trials

Formula for Binomial Distribution in excel

=BINOM.DIST(number_s, trials, probability_s, cumulative)

The BINOM.DIST uses the following arguments:

1. **Number_s** (required argument) – This is the number of successes in trials.
2. **Trials** (required argument) – This is the number of independent trials. It must be greater than or equal to 0.
3. **Probability_s** (required argument) – This is the probability of success in each trial.
4. **Cumulative** (required argument) – This is a logical value that determines the form of the function. It can either be:
    1. TRUE – Uses the cumulative distribution function.
    2. FALSE – Uses the probability mass function.

To learn more, launch our free Excel crash course now!

How to Use the Binomial Distribution Function in Excel?

To understand the uses of the BINOM.DIST function, let us consider an example:

Example

Suppose we are given the following data:

	A	B	C
1			
2		**BINOM.DIST Function**	
3			
4		**Description**	**Data**
5		Number of successes in trials	30
6		Number of independent trials	65
7		Probability of successes in trials	35%
8			

The formula for calculating binomial distribution using the cumulative distribution function is shown below:

SUM =BINOM.DIST(C5,C6,C7,TRUE)

	A	B	C	D	E
1					
2		**BINOM.DIST Function**			
3					
4		**Description**	**Data**		
5		Number of successes in trials	30		
6		Number of independent trials	65		
7		Probability of successes in trials	35%		
8					
9					
10					
11		Binomial distribution using cumulative distribution function	=BINOM.DIST(C5,C6,C7,TRUE)		

We get the result below:

	A	B	C	D
1				
2		**BINOM.DIST Function**		
3				
4		**Description**	**Data**	
5		Number of successes in trials	30	
6		Number of independent trials	65	
7		Probability of successes in trials	35%	
8				
9				
10				
11		Binomial distribution using cumulative distribution function	0.97644	

RANDOM VARIABLE

A **random variable** in statistics is a function that assigns a real value to an outcome in the sample space of a random experiment**.**
For example: if you roll a die, you can assign a number to each possible outcome. Random variables can have specific values or any value in a range.

There are two basic types of random variables,

- Discrete Random Variables
- Continuous Random Variables

Discrete Random Variable

A **Discrete Random Variable** takes on a finite number of values. The probability function associated with it is said to be **PMF**.

PMF(Probability Mass Function)

If X is a discrete random variable then PMF of X is $P(x_i)$, where

P lies between 0 to 1

It is a statistical term that describes the probability distribution of a discrete

random variable.

$$p(x) = P(X = x)$$

The probability of x = the probability(X = one specific x)

Continuous Random Variable takes on an infinite number of values. The probability function associated with it is said to be **PDF (Probability Density Function)**.

PDF (Probability Density Function)

Probability Density Function

$$F(x) = P(a \le x \le b) = \int_a^b f(x)dx \ge 0$$

Then P (X) is said to be a PDF of the distribution.

Q) How to Identify Continuous Distributions? Calculate Continuous distributions.

Q) How to Analyse Continuous probability distributions

Q) How to Analyze data Using Normal distributions.

A) A continuous distribution describes the probabilities of the possible values of a continuous random variable.

A continuous random variable is a random variable with a set of possible values (known as the range) that is infinite and uncountable.

Probabilities of continuous random variables (X) are defined as the area under the curve of its PDF. Thus, only ranges of values can have a nonzero probability. The probability that a continuous random variable equals some value is always zero.

Identify Continuous Distribution

- A probability distribution in which the random variable X can take on any value (is continuous).
- There are infinite values that X could assume.
- The probability of X taking on any one specific value is zero.
- Continuous probability distributions are represented by a **Probability Density Function (PDF)**, which describes the shape of the distribution.

PDF (Probability Density Function)

Probability Density Function

$$F(x) = P(a \le x \le b) = \int_a^b f(x)dx \ge 0$$

- Then P (X) is said to be a PDF of the distribution
- Examples of continuous variables are speed, sound, height, weight, blood pressure, cholesterol levels, etc.

Continuous Probability Distributions

These distributions model the probabilities of random variables that can have any possible outcome.

For example, the possible values for the random variable X that represents weights of citizens in a town which can have any value like 34.5, 47.7, etc.,

Examples: Normal, Student's T, Chi-square, Exponential, etc.,

NORMAL DISTRIBUTION

A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off

symmetrically toward either extreme. The middle of the range is also known as the *mean* of the distribution.

The normal distribution is also known as a *Gaussian distribution* or *probability bell curve*. It is symmetric about the mean and indicates that values near the mean occur more frequently than the values that are farther away from the mean.

- The standard normal distribution has two parameters:**The mean and The standard deviation.**
- The standard normal distribution is a special case of the normal distribution where $\mu=0,\sigma 2=1$. If is often essential to normalize data prior to the analysis.
- In a normal distribution, mean (average), median (midpoint), and mode (most frequent observation) are equal.
- These values represent the peak or highest point.
- The distribution then falls symmetrically around the mean, the width of which is defined by the standard deviation.

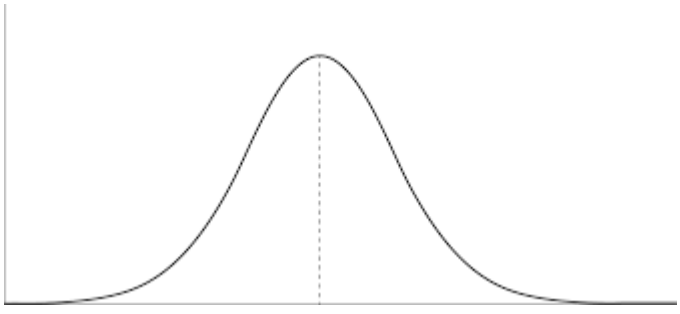
MATHEMTICAL FORMULA :

Normal Distribution Formula

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$\mu =$ mean of x
$\sigma =$ standard deviation of x
$\pi \approx 3.14159 \ldots$
$e \approx 2.71828 \ldots$

GRAPHICAL REPRESENTATION:



The normal distribution describes a symmetrical plot of data around its mean value, where the width of the curve is defined by the standard deviation. It is visually depicted as the "bell curve."

The normal distribution describes a symmetrical plot of data around its mean value, where the width of the curve is defined by the standard deviation.

Calculate Normal Distribution in Excel?

A) Normal distribution in Excel is a way to analyse data for a specific variable. This variable is a representation of the mean of a set of values. The mean is usually an expected value based on your data. The outcome of a normal distribution can help you determine how close your expected variable is to the actual value of the distribution. To find the normal distribution in Excel, you use the NORM.DIST function. The NORM.DIST function has a set of rules known as syntax you follow for it to operate correctly.

The syntax for NORM.DIST is:

=NORM.DIST(x, mean, standard_dev, cumulative)

Where:

- "=" tells Excel you want it to perform a function.
- "NORM.DIST" is the function you want Excel to perform.
- "x" is a required component of the function and represents the value for which you want to find the distribution.
- "mean" is a required component of the function and represents the arithmetic average of the distribution.
- "Standard_dev" is a required component of the function and represents the standard deviation of the distribution.
- "Cumulative" is a required component of the function and represents a logical function of the distribution where TRUE outputs the normal distribution function and FALSE outputs the probability density function.

Example 1:

x = 42, mean = 40 and standard deviation = 1.

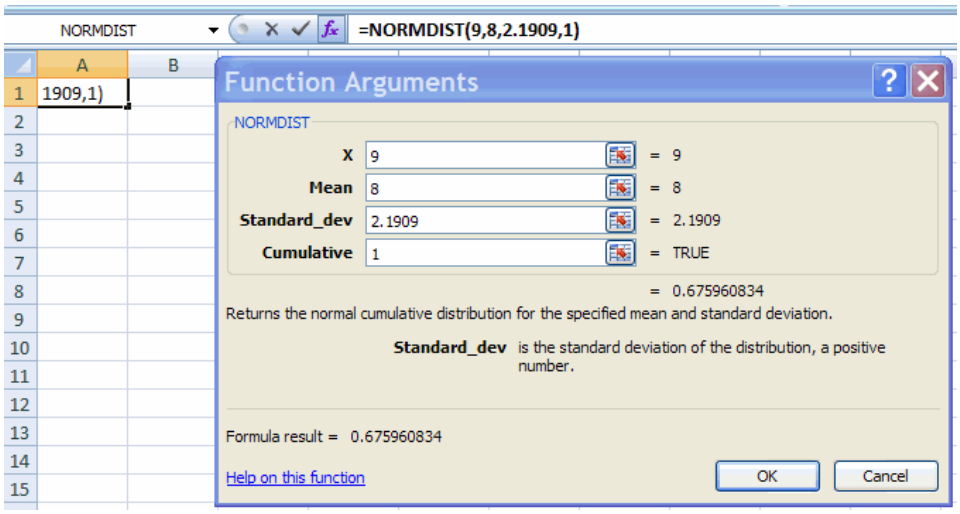
Because he has all the necessary values,

- Type the NORM.DIST function in a blank cell in his Excel spreadsheet.
- Then, he fills it so it reads as =NORM.DIST(42, 40, 1, TRUE).
- When he presses "Enter" on his keyboard,
- Excel outputs the value 0.97725, which is his normal distribution

Example 2:

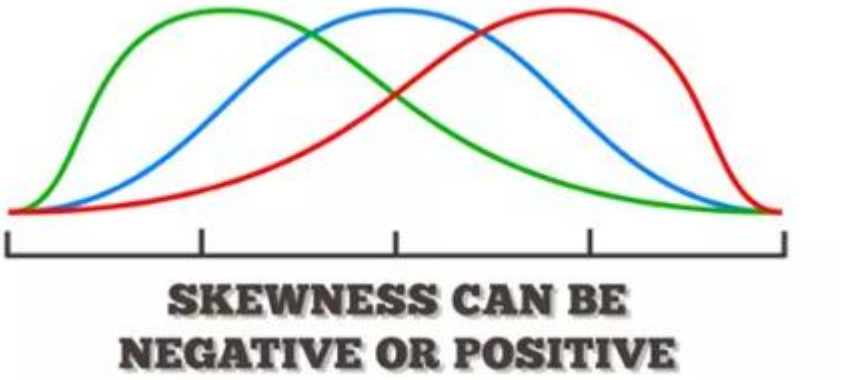
IF VALUES X = 9, mean = 8 , SD = 2.19 Commulative = 1

- Select any cell type = NORMDIST(9,8,2.19,1)
- Press enters



Q) Analyze skewness in Excel .

A) Skewness measures the degree of symmetry of a distribution. The normal distribution is symmetric and has a skewness of zero. If the distribution of a data set instead has a skewness less than zero, or negative skewness (left-skewness), then the left tail of the distribution is longer than the right tail; positive skewness (right-skewness) implies that the right tail of the distribution is longer than the left.



The Skew Function or Data Analysis:

Installing the Toolpak while working on the statistical project gives access to many other functions used in stats.

Skew Function

- Enter the data in Columns.

- Choose a blank cell from the worksheet.
- Add =AKEW(xx:yy)”
- In this formula, xx:yy is the cell location of data.

Data Analysis

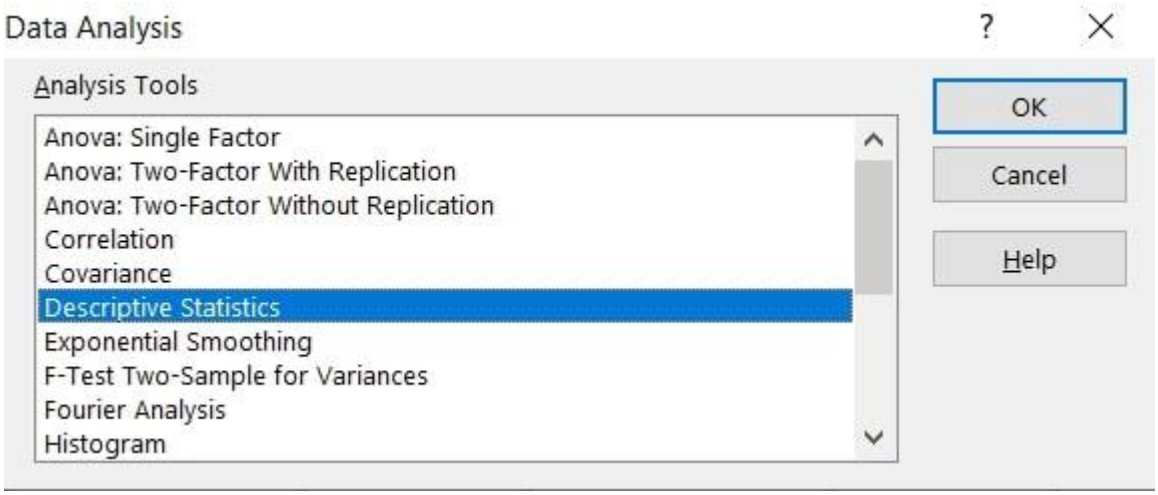
- Click on the Data tab and choose “Data Analysis”.
- Select Descriptive Statistics and press OK.
- Simply click on the input Range box and put the data location.
- Click on the radio button for rows and columns.
- Choose the Labels in the first-row box when the data has column headers.
- Open the Descriptive Statistics.
- Choose a location for the output.
- Press OK.

How to Calculate Skewness with Data Analysis ToolPak Add-In?

To use this method, you need to install the Data Analysis ToolPal first. If you already have installed then follow the steps given below:

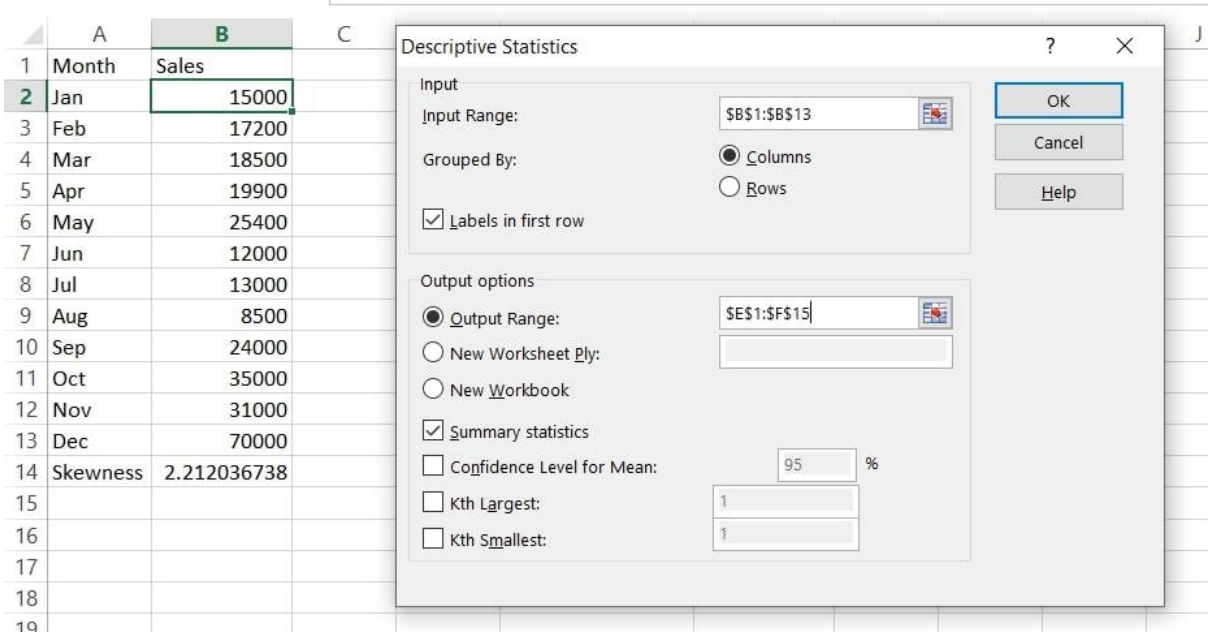
- Open the Data tab from the Ribbon.
- Click on the Data Analysis button.

- Click on the Descriptive Statistics option from the new window that appears.
- Click OK.



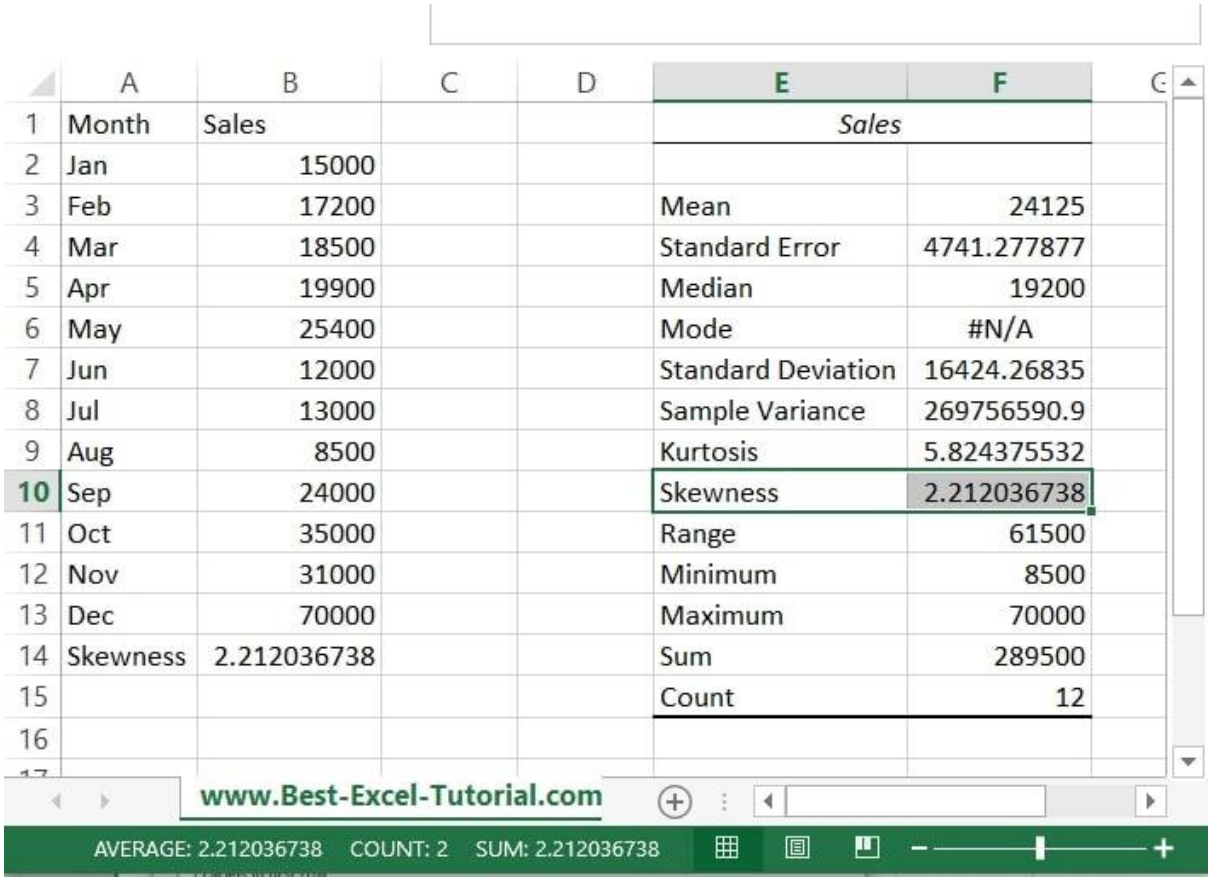
- You have to add parameters:

1. Input Range: It is the range of your data.
2. Labels in the first row: It is the labels in the data table.
3. Output Range: It lets you choose where you need to get the skewness calculated.
4. Summary Statistics: It needs you to check properly.



	A	B
1	Month	Sales
2	Jan	15000
3	Feb	17200
4	Mar	18500
5	Apr	19900
6	May	25400
7	Jun	12000
8	Jul	13000
9	Aug	8500
10	Sep	24000
11	Oct	35000
12	Nov	31000
13	Dec	70000
14	Skewness	2.212036738

In row 10, you can see the skewness calculation and the result is not different than the first example.



	A	B	C	D	E	F
1	Month	Sales			*Sales*	
2	Jan	15000				
3	Feb	17200			Mean	24125
4	Mar	18500			Standard Error	4741.277877
5	Apr	19900			Median	19200
6	May	25400			Mode	#N/A
7	Jun	12000			Standard Deviation	16424.26835
8	Jul	13000			Sample Variance	269756590.9
9	Aug	8500			Kurtosis	5.824375532
10	Sep	24000			Skewness	2.212036738
11	Oct	35000			Range	61500
12	Nov	31000			Minimum	8500
13	Dec	70000			Maximum	70000
14	Skewness	2.212036738			Sum	289500
15					Count	12
16						

Advantages

Let us discuss the advantages of negative and **positive skewness** through the discussion below.

- Skewness is better for measuring the performance of investment returns.
- The investor uses it when analyzing the data set as it considers the extreme of the distribution.
- It is a widely used tool in statistics as it helps understand how much data is asymmetry from the normal distribution.

Disadvantages

Despite the various advantages of the **skewness distribution** as discussed above, there are a few disadvantages that make using this statistical tool a hassle. Let us cover all bases of the concept through the points below.

- Skewness ranges from negative to positive infinity. Sometimes, it is difficult for an investor to predict the trend in the data set.

- An analyst is forecasting the future performance of an asset using the financial model, which usually assumes that data is normally distributed. But, if the data distribution skews, this model will not reflect the actual result in its assumption.

Introduction to R- Programming

Introduction:

- R is a **programming language** and software environment for **statistical analysis**,**graphics representation and reporting**.R was created by **Ross Ihaka** and **Robert Gentleman** at the University of Auckland, NewZealand, and is currently developed by the R Development Core Team.
- R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.
- This programming language was named R,based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka),and partly a play on the name of theBell Labs Language S.
- R is the most popular data analytics tool as it is open-source, flexible, offers multiple packages and has a huge community.

Q) What is Business Analytics?

A) Business analytics is a process of examining large sets of data and achieving hidden patterns, correlations and other insights. It basically helps you understand all the data that you have gathered, be it organizational data, market or product research data or any other kind of data. It becomes easy for you to make better decisions, better products, better marketing strategies etc.

But you can't expect people in the chain above to always understand the raw data that you are providing them after analytics. So to overcome this gap, we have a concept of *data visualization*.

Data visualization: Data visualization is a visual access to huge amounts of data that you have generated after analytics. The human mind processes visual images and visual graphics are better than compare to raw data. It's always easy for us to understand a pie chart or a bar graph compare to raw numbers. Now you may be wondering how you can achieve this data visualization from the data you have already analyzed.

There are various tools available in the market for Data Visualization: R, Power BI, Spark, Qlikview etc.

Q) What is Business Analytics and Explain the Role of R in Business Analytics?

A) Business analytics is a process of examining large sets of data and achieving hidden patterns, correlations and other insights.

let us see some of the problems and their solutions in R in multiple domains.

Banking:

Large amount of customer data is generated every day in Banks. While dealing with millions of customers on regular basis, it becomes hard to track their mortgages.

Solution:

R builds a custom model that maintains the loans provided to every individual customer which helps us to decide the amount to be paid by the customer over time.

Insurance:

Insurance extensively depends on forecasting. It is difficult to decide which policy to accept or reject.

Solution:

By using the continuous credit report as input, we can create a model in R that will not only assess risk appetite but also make a predictive forecast as well.

Healthcare:

Every year millions of people are admitted in hospital and billions are spent annually just in the admission process.

Solution:

Given the patient history and medical history, a predictive model can be built to identify who is at risk for hospitalization and to what extent the medical equipment should be scaled.

Q) What is R Programming? Explain Evolution, Uses, Applications and Features of R?

A)R is a **programming language** and software environment for **statistical analysis**, **graphics representation and reporting**. R was created by **Ross Ihaka** and **Robert Gentleman.** R is the most popular data analytics tool as it is open-source, flexible, offers multiple packages and has a huge community

Evolution of R

- R is an implementation of S programming language which was created by John Chambers at Bell Labs.
- R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University of Auckland in Auckland, New Zealand.
- R made its first public appearance in 1993.
- A large group of individuals has contributed to R by sending code and bug reports.Since mid-1997 there has been a core group (the "R Core Team") who can modify the R source code archive.
- In the year 2000 R 1.0.0 released.
- R 3.0.0 was released in 2013.

USES OF R

- R is a programming and statistical language.
- R is used for data Analysis and Visualization.
- R is simple and easy to learn, read and write.
- R is an example of a FLOSS (Free Libre and Open Source Software) where one can freely distribute copies of this software, read its source code, modify it, etc.

APPLICATIONS OF R

- The Consumer Financial Protection Bureau uses R for data analysis
- Statisticians at John Deere use R for time series modeling and geospatial analysis in a reliable and reproducible way.
- Bank of America uses R for reporting.
- R is part of technology stack behind Foursquare's famed recommendation engine.
- ANZ, the fourth largest bank in Australia, using R for credit risk analysis.
- Google uses R to predict Economic Activity.
- Mozilla, the foundation responsible for the Firefox web browser, uses R to visualize Web activity.

Features of R:

- R supports procedural programming with functions and **object-oriented programming** with generic functions.

- **Packages** are part of R programming. Hence, they are useful in collecting sets of **R functions** into a single unit.
- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility,
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- It provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.
- Rs programming features include database input, exporting data, viewing data, variable labels, missing data, etc.
- R is an interpreted language. So we can access it through command line interpreter.
- R supports **matrix** arithmetic.
- R, SAS, and SPSS are three statistical languages. Of these three statistical languages, R is the only an open source.

As a conclusion, R is world's most widely used statistics programming language. It is a good choice of data scientists and supported by a vibrant and talented community of contributors.

The prominent editors available for R programming language are:

- **RGUI**(R graphical user interface) - Rstudio – Studio R offers a richer editing environment than RGUI and makes some common tasks easier and more fun.
- **RStudio -** RStudio is an integrated development environment (IDE) for R language. RStudio is a code editor and development environment, with some nice features that make code development in R easy and fun.
- **R Command Prompt**

Q) Write a procedure for Install R and R-Studio for Windows

1. To install R, go to **cran.r-project.org**
2. Depending on your operating system, click Download R for (your operating system).

3. Click on install R for the first time.
4. Click Download R for Windows. Open the downloaded file.
5. Select the language you would like to use during the installation. Then click OK.
6. Click Next.
7. Select where you would like R to be installed. It will default to your Program Files on your C Drive. Click Next.
8. You can then choose which installation you would like.
9. (Optional) If your computer is a 64-bit, you can choose the 64-bit User Installation. Then click Next.
10. Then specify if you want to customized your startup or just use the defaults. Then click Next.
11. Then you can choose the folder that you want R to be saved within or the default if the R folder that was created. Once you have finished, click Next.
12. You can then select additional shortcuts if you would like. Click Next.
13. Click Finish.

Once you have R environment setup, then it's easy to start your R command prompt by just clicking on R Software icon. This will launch R interpreter and you will get a prompt > where you can start typing your programs or commands.

```
>x=6
>print(x)
```

14. Next, download RStudio. Go to **www.rstudio.com**
15. Click Download RStudio.
6. Click Download under RStudio Desktop- Open Source License.
17. Click on the operating system that you are working with.
18. The RStudio installation wizard will pop-up. Click Next and go through the installation steps.
19. Congratulations! You have now installed R and RStudio.

Q) Sample Program Using R

Usually, you will do your programming by writing your programs in script files and then you execute those scripts at your command prompt with the help of R interpreter called Rscript. So let's start with writing following code in a text file called test.R as under –

```
# My first program in R Programming
myString<- "Hello, World!" print ( myString)
```

Save the above code in a file **test.R**. Execute by opening that script in R editor, select all (**Ctrl +A**) and click on **run line** or selection (**Ctrl+R)** option in Edit menu of R console.

When we run the above program, it produces the following output.

[1] "Hello, World!"

Q) Explain Variables and Operators in R Programming?

Variables: A variable provides us with named storage that our programs can manipulate. A variable in R can store an atomic vector, group of atomic vectors or a combination of many R- objects. A valid variable name consists of letters, numbers and the dot or underline characters. The variable name starts with a letter or the dot not followed by a number.

Variable Name	Validity	Reason
var_name2.	valid	Has letters, numbers, dot and underscore
var_name%	Invalid	Has the character '%'. Only dot(.) and underscore allowed.
2var_name	invalid	Starts with a number
.var_namevar.name’	valid	Can start with a dot(.) but the dot(.)should not be followed by a number.
.2var_name	invalid	The starting dot is followed by a number making it invalid.
_var_name	invalid	Starts with _ which is not valid

Variable Assignment

The variables can be assigned values in 3 ways

1. Assignment using equal operator.

```
var.1 = c(0,1,2,3)
```

2. Assignment using leftward operator.

```
var.2 <- c("learn","R")
```

3. Assignment using rightward operator.

```
c(TRUE,1) -> var.3
```

Printing a Variable Value: The values of the variables can be printed using print() or cat()function. The cat() function combines multiple items into a continuous print output.

```
print(var.1) cat ("var.1 is ", var.1 ,"\n") cat ("var.2 is ", var.2 ,"\n") cat ("var.3 is ", var.3 ,"\n")
```

output

```
[1] 0 1 2 3 var.1 is  0 1 2 3  var.2 is  learn R  var.3 is  1 1
```

Operators:

An operator is a symbol that tells the compiler to perform specific mathematical or logical manipulations. R language is rich in built-in operators and provides following types of operators.

Types of Operators

We have the following types of operators in R programming –

- Arithmetic Operators
- Relational Operators
- Logical Operators
- Assignment Operators
- Miscellaneous Operators

Arithmetic Operators:

Following table shows the arithmetic operators supported by R language. The operators act on each element of the vector

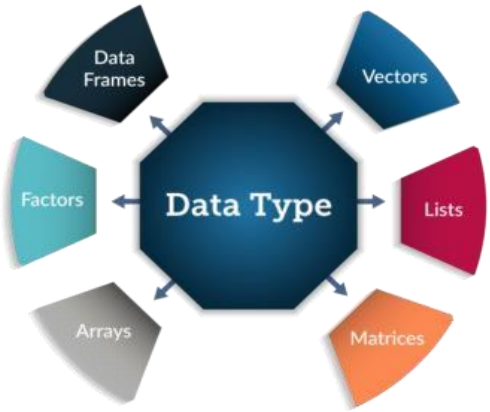
Operator	Description	Example
+	Adds two vectors	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v+t) it produces the following result – [1] 10.0 8.5 10.0
–	Subtracts second vector from the first	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v-t) it produces the following result – [1] -6.0 2.5 2.0
*	Multiplies both vectors	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v*t) it produces the following result – [1] 16.0 16.5 24.0
/	Divide the first vector with the second	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v/t) When we execute the above code, it produces the following result – [1] 0.250000 1.833333 1.500000
%%	Give the remainder of the first vector with the second	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v%%t) it produces the following result – [1] 2.0 2.5 2.0

%/%	The result of division of first vector with second (quotient)	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v%/%t) it produces the following result – [1] 0 1 1
^	The first vector raised to the exponent of second vector	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v^t) it produces the following result – [1] 256.000 166.375 1296.000

Data Objects in R:

Data types are used to store information. In R, we do not need to declare a variable as some data type. The variables are assigned with R-Objects and the data type of the R-object becomes the data type of the variable.There are mainly six data types present in R:

1. **Vectors**
2. **Lists**
3. **Matrices**
4. **Arrays**
5. **Factors**
6. **Data Frames**



Scalar variable :A scalar is a single number. The following code creates. a scalar variable with the numeric value 5: x = 5. Vector variable A vector is a sequence of numbers.

Q) Explain Vectors and Data Frames in R

A)

1.Vector:

In R, a sequence of elements which share the same data type is known as vector. A vector supports logical, integer, double, character, complex, or raw data type. The elements which are contained in vector known as **components** of the vector. We can check the type of vector with the help of the **typeof()** function.

How to create a vector in R?

In R, we use **c() function** to create a vector. This function returns a one-dimensional array or simply vector. The c() function is a generic function which combines its argument.

There are various other ways to create a vector in R, which are as follows:

1) Using the colon(:) operator

We can create a vector with the help of the colon operator. There is the following syntax to use colon operator:

syntax
z<-x:y

This operator creates a vector with elements from x to y and assigns it to z.

Example:

```
1.  a<-4:-10
2.  a
```

Output

```
[1]  4  3  2  1  0 -1 -2 -3 -4 -5 -6 -7 -8 -9 -10
```

2) Using the seq() function

In R, we can create a vector with the help of the seq() function. A sequence function creates a sequence of elements as a vector. The seq() function is used in two ways, i.e., by setting step size with ?by' parameter or specifying the length of the vector with the 'length.out' feature.

```
seq_vec<-seq(1,4, by=0.5)

seq_vec
```

Output

```
[1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0
```

There are 5 Atomic vectors, also termed as five classes of vectors.

Data Frames in R

Data Frames: A data frame is a table or a two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column. Data frames are tabular data objects. Unlike a matrix in data frame each column can contain different modes of data. The first column can be numeric while the second column can be character and third column can be logical. It is a list of vectors of equal length. Data Frames are created using the **data.frame()** function.

```
# Create the data frame.
BMI <-data.frame(gender=c("Male", "Male","Female"),
height = c(152, 171.5, 165),
weight = c(81,93, 78),
                Age = c(42,38,26)
)
print(BMI)
```

OUTPUT

```
  gender height weight Age
1   Male  152.0     81  42
2   Male  171.5     93  38
3 Female  165.0     78  26
```

Q) Explain How to Read another file into R

One of the most important features we need to be able to do in R is import existing data, whether it be .txt files, .csv files, or even .xls (Excel files).

Reading Tabular Data Files or .txt files:

If you have a **.txt or a tab-delimited text file**, or a file with table like structure then you can easily import it by using the basic R function **read.table().**

Syntax:

```
read.table(file, header = FALSE, sep = "", quote = "\"", dec = ".", row.names, col.names, ..)
```

Example 1: A data table can resides in a text file. The cells inside the table are separated by blank characters. Here is an example of a table with 4 rows and 3 columns.

```
100  a1  b1
200  a2  b2
300  a3  b3
400  a4  b4
```

```
>mydata = read.table("mydata.txt")
 # read text file
>mydata
 # print data frame
     V1 V2 V3
1   100 a1 b1
2   200 a2 b2
3   300 a3 b3
4   400 a4 b4
```

Reading CSV Files:

The csv file is a text file in which the values in the columns are separated by a comma. While R can read excel .xls and .xlsx files these file types often cause problems.

If you need to read in a csv with R the best way to do is by using **read.csv()** function.

Syntax:

```
read.csv (file, header = TRUE, sep = ",", quote = "\"", dec = ".", ......)
```

Example1: Let's consider the following data present in the file named **input.csv**.

```
data <- read.csv("input.csv")
print(data)
```

```
O/P:  id,  name,   salary,  start_date,    dept
1       1   Rick    623.30   2012-01-01     IT
2       2   Dan     515.20   2013-09-23     Operations
3       3   Michelle 611.00   2014-11-15     IT
4       4   Ryan    729.00   2014-05-11     HR
5       NA Gary     843.25   2015-03-27     Finance
6       6   Nina    578.00   2013-05-21     IT
```

Importing Data from Excel Files:

Microsoft Excel is the most widely used spreadsheet program which stores data in the .xls or .xlsx format. R can read directly from these files using some excel specific packages. Few such packages are - XLConnect, xlsx, gdata etc. We will be using xlsx package.

- **Importing and loading "xlsx" package:**

You can use the following command in the R console to install the "xlsx" package. It may ask to install some additional packages on which this package is dependent.

```
install.packages("xlsx")
```

- **Load the imported package into R workspace**

```
# Load the library into R workspace.

library("xlsx")
```

- **Reading the Excel File**

The input.xlsx is read by using the**read.xlsx()** function as shown below. The result is stored as a data frame in the R environment.

```
# Read the first worksheet in the file input.xlsx.
```

```
data<- read.xlsx("input.xlsx", sheetIndex = 1)

print(data)
```

- We can read an excel file by selecting sheet index or sheet name.

```
data<- read.xlsx("input.xlsx", sheetIndex = 1) or

data<- read.xlsx("input.xlsx", sheetName = “sheet1”)
```

Q) Explain How to Apply Mean/Median/Standard Deviation in R

A) Mean

It is calculated by taking the sum of the values and dividing with the number of values in a data series.

The function **mean()** is used to calculate this in R.

Syntax **mean(x, trim = 0, na.rm = FALSE, ...)**

Following is the description of the parameters used –

- **x** is the input vector.
- **trim** is used to drop some observations from both end of the sorted vector.
- **na.rm** is used to remove the missing values from the input vector.

Example

```
# Create a vector.
x <-c(12,7,3,4.2,18,2,54,-21,8,-5)

# Find Mean.
result.mean<- mean(x)
print(result.mean)
```

output
[1] 8.22

Median

The middle most value in a data series is called the median. The **median()** function is used in R to calculate this value.

Syntax

median(x, na.rm = FALSE)

Following is the description of the parameters used –

- **x** is the input vector.
- **na.rm** is used to remove the missing values from the input vector.

Example

```
# Create the vector.
x <-c(12,7,3,4.2,18,2,54,-21,8,-5)

# Find the median.
median.result<- median(x)
print(median.result)
```

output

[1] 5.6

Standard deviation

- **'Standard deviation is the measure of the dispersion of the values'.**
- The higher the standard deviation, the wider the spread of values.
- The lower the standard deviation, the narrower the spread of values.
- In simple words the formula is defined as - **Standard deviation is the square root of the 'variance'.**

The standard deviation is computed with the sd() function.

Let us create a list 'x' and add some value to it. Then we can find the standard deviation of those values in the list.

```
x <-c(34,56,87,65,34,56,89)#creates list 'x' with some values in it.

sd(x)#calculates the standard deviation of the values in the list 'x'
```

Output —> 22.28175

Finding the Standard deviation of the values stored in a CSV file

Here importing a CSV file to find the standard deviation in R for the values which are stored in that file.

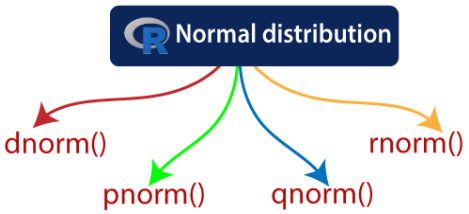
```
readfile<- read.csv('testdata1.csv')#reading a csv file
data2 <-readfile$Values#getting values stored in the header 'Values'
sd(data2)#calculates the standard deviation
```

Q) Explain Normal Distribution and Poisson Distribution Functions in R

A)

R Normal Distribution: If we plot a graph with the value of the variable in the horizontal axis and counting the values in the vertical axis, then we get a bell shape curve. The curve center represents the mean of the data set. In the graph, fifty percent of the value is located to the left of the mean. And the other fifty percent to the right of the graph. This is referred to as the normal distribution.

R allows us to generate normal distribution by providing the following functions:



These function can have the following parameters:

S.No	Parameter	Description
1.	X	It is a vector of numbers.
2.	P	It is a vector of probabilities.
3.	N	It is a vector of observations.
4.	Mean	It is the mean value of the sample data whose default value is zero.
5.	Sd	It is the standard deviation whose default value is 1.

Let's start understanding how these functions are used with the help of the examples.

dnorm():Density

The dnorm() function of R calculates the height of the probability distribution at each point for a given mean and standard deviation. The probability density of the normal distribution is:

Syntax

dnorm(x, mean ,sd)

Parameters:

- *x: vector of quantiles.*
- *mean: vector of means.*
- *sd: vector standard deviation.*

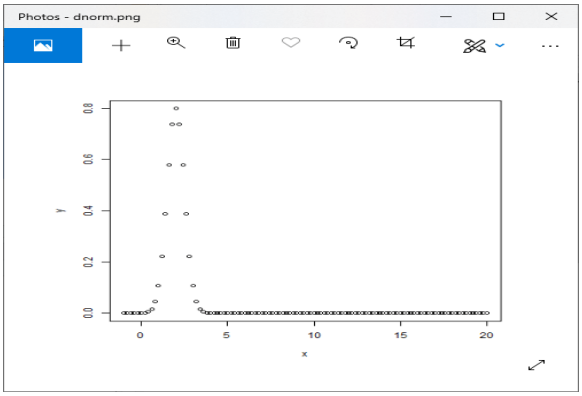
Example

dnorm(x=1, mean=0, sd=1)

Program # Creating a sequence of numbers between -1 and 20 incrementing by 0.2.

```
x <- seq(-1, 20, by = .2)
# Choosing the mean as 2.0 and standard deviation as 0.5.
y <- dnorm(x, mean = 2.0, sd = 0.5)
```

Output:



POISSON DISTRIBUTION IN R

The Poisson distribution represents the probability of a provided number of cases happening in a set period of space or time if these cases happen with an identified constant mean rate (free of the period since the ultimate event). Poisson distribution has been named after Siméon Denis Poisson(French Mathematician).
Many probability distributions can be easily implemented in R language with the help of R's inbuilt functions.
There are four Poisson functions available in R:

- dpois
- ppois
- qpois
- rpois

dpois()

This function is used for illustration of Poisson density in an R plot. The function dpois() calculates the probability of a random variable that is available within a certain range.

Syntax:

dpois(k,λ,log)

where,

K: *number of successful events happened in an interval*

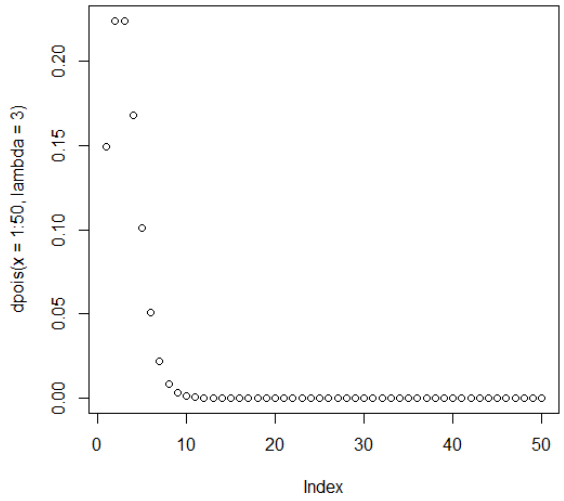
mean per interval

log: *If TRUE then the function returns probability in form of log*

Example

```
plot(dpois(x=1:50,lambda=3))
```

Output



Q) Explain Bayes Theorem in R

Bayes' theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of "causes"

As a result, we now have an ideal Bayes Classifier against which all other classifiers are judged.

The following are the three most common applications of the Bayes theorem

1) Bayes' Naive Classifiers

2) Decision Surfaces and Discriminant Functions

3) Estimation of Bayesian Parameters

how to use the Bayes theorem in R.

Example: Assume that there is a 30% chance of rainfall in a day. Assume that the probability that I walk outside is 50% and that the probability that I walk on a rainy day is 10%. What are the chances that it will be a rainy day given I walk outside?

This problem can be solved by using the Bayes Theorem. Wrapping this data around on the head is quite tough. Let's change it into probability notation.

Let's assume:

The probability of raining in a day P(R) = 30% = 0.30

The probability that I walk outside P(W) = 50% = 0.50

The Probability that I walk in a (or given) rainy day P(W|R) = 10% = 0.10

The probability of rain when (or given) I walk P(R|W) = ?

Bayes' Theorem in R

```
BayesTheorem = function(PA, PB, PA_PB) {
  PA_PB = (PA * PA_PB )/ PB
  return(PA_PB)
}

PRain = 0.30
PWalk = 0.50
PWalkGRain = 0.10

BayesTheorem(PRain, PWalk, PWalkGRain)
```

This should give a value of P(R|W) = 0.06

Based on certain conditions, the probability of it Rainy day is 0.06.

Q) Explain Scatter plots and Box plot

Scatter Plots

A "scatter plot" is a type of plot used to display the relationship between two numerical variables, and plots one dot for each observation.

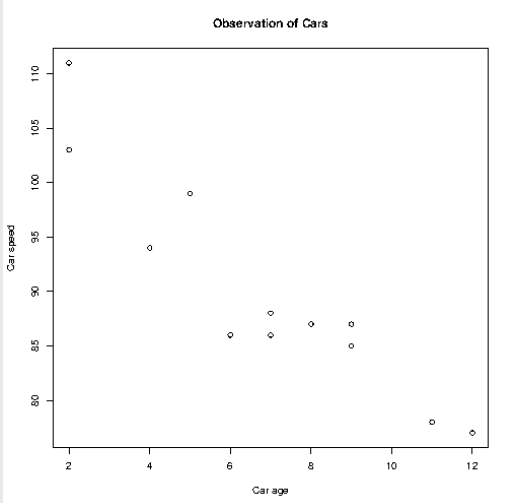
It needs two vectors of same length, one for the x-axis (horizontal) and one for the y-axis (vertical):

Example: The observation of 12 cars passing by.

```
x <- c(5,7,8,7,2,2,9,4,11,12,9,6)
y <- c(99,86,87,88,111,103,87,94,78,77,85,86)

plot(x, y, main="Observation of Cars", xlab="Car age", ylab="Car speed")
```

Result:



The observation in the example above is the result of 12 cars passing by.

The **x-axis** shows how old the car is.

The **y-axis** shows the speed of the car when it passes.

Are there any relationships between the observations?

It seems that the newer the car, the faster it drives, but that could be a coincidence, after all we only registered 12 cars.

BOXPLOTS

Boxplots are a measure of how well distributed is the data in a data set. It divides the data set into three quartiles. This graph represents the minimum, maximum, median, first quartile and third quartile in the data set. It is also useful in comparing the distribution of data across data sets by drawing boxplots for each of them.

Boxplots are created in R by using the **boxplot()** function.

Syntax

boxplot(x, data, notch, varwidth, names, main)

Following is the description of the parameters used –

- **x** is a vector or a formula.
- **data** is the data frame.
- **notch** is a logical value. Set as TRUE to draw a notch.
- **varwidth** is a logical value. Set as true to draw width of the box proportionate to the sample size.
- **names** are the group labels which will be printed under each boxplot.
- **main** is used to give a title to the graph.

Example

We use the data set "mtcars" available in the R environment to create a basic boxplot. Let's look at the columns "mpg" and "cyl" in mtcars.

```
input<-mtcars[,c('mpg','cyl')]
print(head(input))
```

When we execute above code, it produces following result –

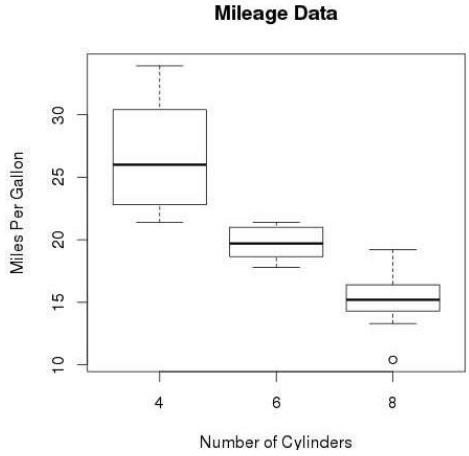
	mpg	cyl
Mazda RX4	21.0	6
Mazda RX4 Wag	21.0	6
Datsun 710	22.8	4
Hornet 4 Drive	21.4	6
Hornet Sportabout	18.7	8
Valiant	18.1	6

Creating the Boxplot

The below script will create a boxplot graph for the relation between mpg (miles per gallon) and cyl (number of cylinders).

```
# Plot the chart.
boxplot(mpg ~cyl, data =mtcars,xlab="Number of Cylinders",
ylab="Miles Per Gallon", main ="Mileage Data")
```

Output



HISTOGRAM

A histogram represents the frequencies of values of a variable bucketed into ranges. Histogram is similar to bar chat but the difference is it groups the values into continuous ranges. Each bar in histogram represents the height of the number of values present in that range.

R creates histogram using **hist()** function. This function takes a vector as an input and uses some more parameters to plot histograms.

Syntax

hist(v,main,xlab,xlim,ylim,breaks,col,border)

Following is the description of the parameters used –

- **v** is a vector containing numeric values used in histogram.
- **main** indicates title of the chart.
- **col** is used to set color of the bars.
- **border** is used to set border color of each bar.
- **xlab** is used to give description of x-axis.
- **xlim** is used to specify the range of values on the x-axis.
- **ylim** is used to specify the range of values on the y-axis.
- **breaks** is used to mention the width of each bar.

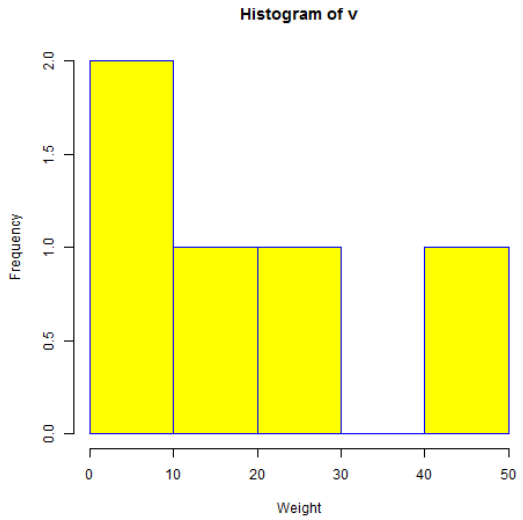
Example

A simple histogram is created using input vector, label, col and border parameters.

```
# Create data for the graph.
v <-  c(9,13,21,8,36,22,12,41,31,33,19)

# Create the histogram.
hist(v,xlab = "Weight",col = "yellow",border = "blue")
```

Output

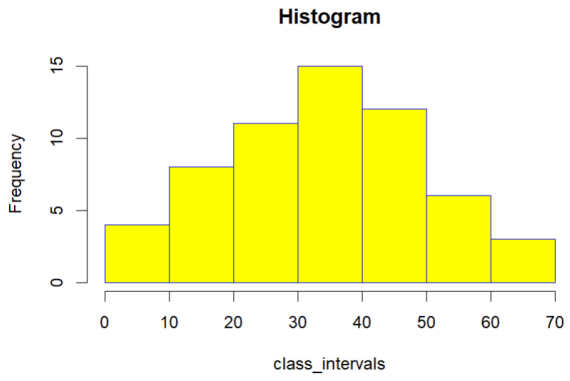


Q) Construct a Histogram for the following data?

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	4	8	11	15	12	6	3

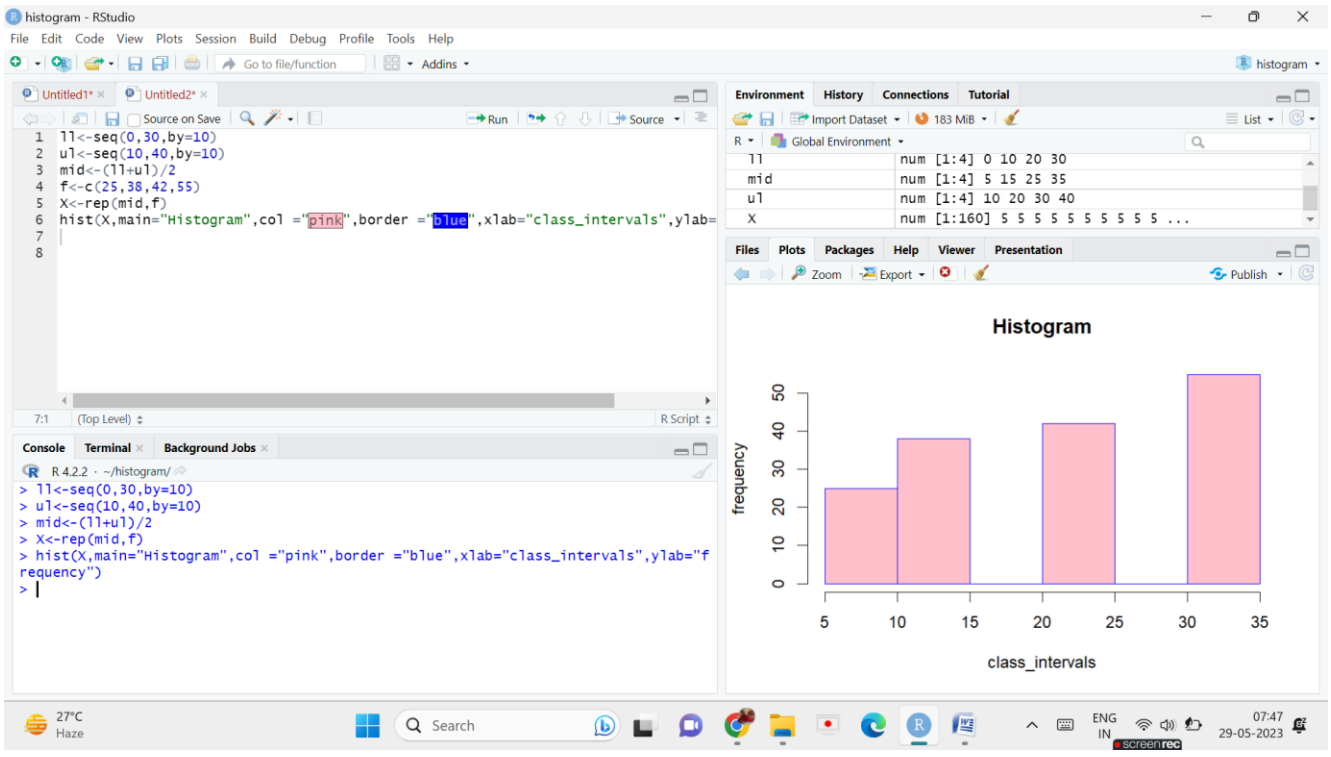
```
> low_lim<-seq(0,60,by=10)
> up_lim<-seq(10,70,by=10)
> mid<-(low_lim+up_lim)/2
> frequency<-c(4,8,11,15,12,6,3)
> X<-rep(mid,frequency)
> hist(X, main = "Histogram",col = "yellow",border="blue",xlab = "cl
ass_intervals",ylab="Frequency" )
```

OUTPUT



Q)Draw Histogram by using 'r' for the following data

Class Interval	0-10	10-20	20-30	30-40
Frequency	25	38	42	55



Q:Explain User Defined Function?

A) A function is just a block of code that you can call and run from any part of your program. They are used to break our code in simple parts and avoid repeatable codes.

You can pass data into functions with the help of parameters and return some other data as a result. You can use the function() reserve keyword to create a function in R.

Syntax

```
Function_name <- function(a,b,...)
{
   ---- Function Body ----
}
```

where 'a' and 'b' are arguments.

And once you define a function, you can call it later as

```
Result <- Function_name(a,b)
```

Ex: Addition of two numbers Using a user defined function

```
MyAddFunction <- function(a,b){
 output = a + b
 result = paste0("Addition of ", a, " and ", b, " gives ", output)
 return(result)
}
```

Try calling this function using different values for a and b as given below:

For a = 3 and b = 5,

MyAddFunction(3,5)

gives result as "Addition of 3 and 5 gives 8"