**Week 4**

**Data engineering pipeline**

**1. Data Collection**

**1.1 Population and Sample Statistic**

**Population:**

* A complete collection of the objects or measurements is called the population.
* Or else everything in the group we want to learn about will be termed as population.
* In statistics population is the entire set of items from which data is drawn in the statistical study. It can be a group of individuals or a set of items.

**Population is the entire group you want to draw conclusions about.**

The population is usually denoted with **N**

1. The number of citizens living in the State of Rajasthan represents a population of the state
2. All the chess players who have FIDE rating represents the population of the chess fraternity of the world
3. the number of planets in the entire universe represents the planet population of the entire universe
4. The types of candies and chocolates are made in India.

population mean is usually denoted by the Greek letter μ

μ (population mean) =

**For example**: let us assume that there are 5 employees in a company, so 5 people is a complete set hence it will represent the population of my company

If we want to find the average age of the company then we simply add their ages and divide it by N which is the number of population

ages = {23,45,12,34,22}

                                                                  μ =

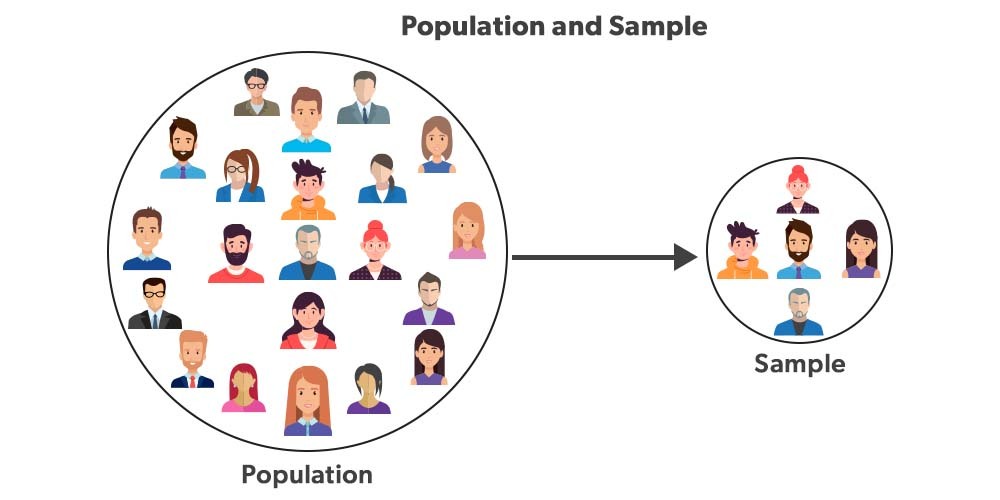
= (23 + 45 + 12 + 34 + 22) / 5

The average age of the company is 27.2 years

* Here the population was quite less so the calculation of the population means was an easy task.
* What if we want to calculate the average height of the Indians, it would be next to an impossible task.
* Even if it is not impossible it would be a difficult task.
* **Since in this case and many others it is impossible to observe the entire statistical population due to the time constraints, constraints on geographical accessibility and the constraints on the researcher resources a researcher would instead observe a sample of the same population in order to attempt to learn something about the population as a whole.**

**Sample:**

* A sample represents a group of the interest of the population which we will use to represent the data.
* The sample is an unbiased subset of the population in which we represent the whole data.
* A sample is a group of the elements actually participating in the survey or study.
* A sample is the representation of the manageable size.
* Samples are collected and stats are calculated from the sample so one can make interferences or extrapolations from the sample. This process of collecting info from the sample is called **sampling.**
* The sample is denoted by **n**



1. 500 people from a total population of the Rajasthan state will be considered as a sample
2. 150 total chess players from all total number of chess players will be considered as a sample

*Sample mean is denoted by x –*

x̄ *(****sample mean****) =*=

**for example:** The total number of users of a is the population and, all student accounts of the website is a sample.

**2. Types of Data**

### 2.1 Data Type

### 2.1.1 Type 1

### a) Cross-sectional Data

Cross-section data is collected in a single time period and is characterized by individual units - people, companies, countries, etc. Some examples include:

* Student grades at the end of the current semester;
* Household data of the previous year - expenditure on food, unemployment, income, etc.
* Car data - average speed, horsepower, colour, etc.
* **With cross-sectional data the ordering of the data does not matter**. In other words, we can order the data by ascending, descending or even randomized order and this will not affect out modelling results.
* Cross-sectional data or cross-section of a population is obtained by taking observations from multiple individuals at the same point in time.
* Cross-sectional data can comprise of observations taken at different points in time, however, in such cases time itself does not play any significant role in the analysis.
* Exam scores of high school students in a particular year is an example of cross-sectional data.
* Gross domestic product of countries in a given year is another example of cross-sectional data.
* However, we need to note that, in case of exam scores of students and GDP of countries, all the observations have been taken in a single year and this makes the two datasets cross-sectional.
* In essence, the cross-sectional data represents a snapshot at a given instance of time in both the cases.

### b) Time Series Data

Data collected at a number of specific points in time is called time series data. Such examples include stock prices, interest rates, exchange rates as well as product prices, GDP, etc.

* Time series data can be observed at many different frequencies (hourly, daily, weekly, monthly, quarterly, annually etc)
* **Unlike cross-sectional data, the ordering of the data is important in time-series data.**
* Each point represents the values at specific points in time.
* As such, time series data are typically presented in chronological order.
* Changing the order of the data ignores the time-dimensionality of the data.

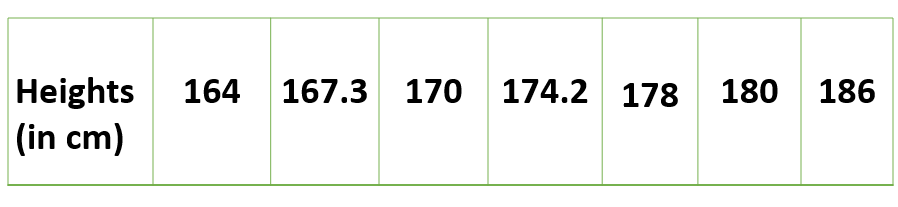
**Example:**

* If we consider only one country, and take a look at its military expenses and central government debt for a span of 10 years from 2001 to 2010, that would get two time series - one about the military expenditure and the other about debt of central government debt.
* Therefore, in essence, a time series is made up of quantitative observations on one or more measurable characteristics of an individual entity and taken at multiple points in time.
* Time series data is typically characterized by several interesting internal structures such as trend, seasonality, stationarity, autocorrelation, and so on.

**2.1.2 Type 2**

**a) Univariate data**

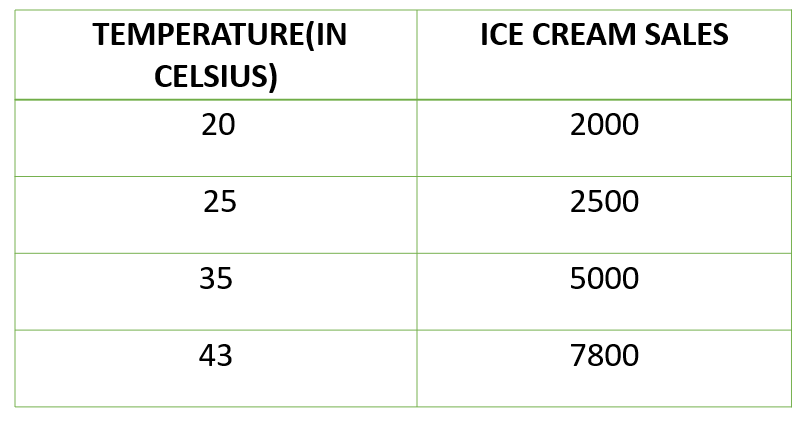
* This type of data consists of **only one variable**.
* The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes.
* It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
* The example of a univariate data can be height.



* Suppose that the heights of seven students of a class is recorded, there is only one variable that is height and it is not dealing with any cause or relationship.
* The description of patterns found in this type of data can be made by drawing conclusions using central tendency measures (mean, median and mode), dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation) and by using frequency distribution tables, histograms, pie charts, frequency polygon and bar charts.

**b) Bivariate data**

* This type of data involves **two different variables**.
* The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.
* Example of bivariate data can be temperature and ice cream sales in summer season.



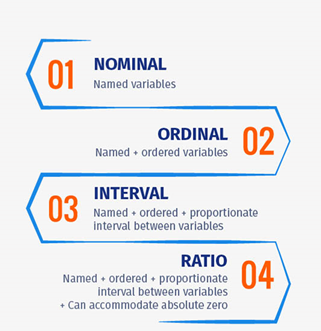
* Suppose the temperature and ice cream sales are the two variables of a bivariate data.
* Here, the relationship is visible from the table that temperature and sales are directly proportional to each other and thus related because as the temperature increases, the sales also increase.
* Thus, bivariate data analysis involves comparisons, relationships, causes and explanations.
* These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent.

**c) Multivariate data**

* When the data involves **three or more variables**, it is categorized under multivariate.
* Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.
* It is similar to bivariate but contains more than one dependent variable.
* The ways to perform analysis on this data depends on the goals to be achieved.
* Some of the techniques are regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA)

**2.2 Variable Types**

* A variable is a characteristic that can be measured and that can assume different values.
* A variable can occur in any form, such as trait, factor or a statement that will constantly be changing according to the changes in the applied environment.
* Height, age, income, province or country of birth, grades obtained at school and type of housing are all examples of variables.
* In the 1940s, Stanley Smith Stevens introduced four scales of measurement:
  + nominal
  + ordinal
  + interval
  + ratio
* These are still widely used today as a way to describe the characteristics of a variable.
* Knowing the scale of measurement for a variable is an important aspect in choosing the right statistical analysis.

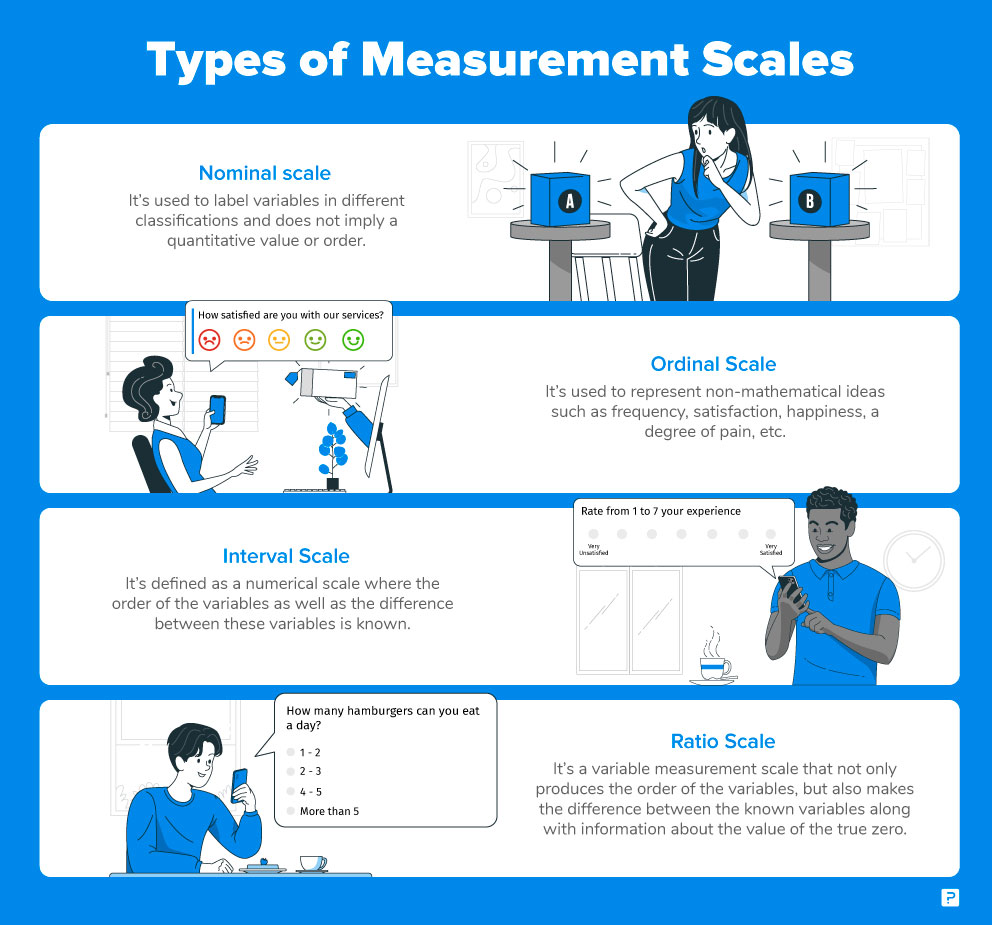


## Levels of Measurement in Statistics

* To perform statistical analysis of data, it is important to first understand variables and what should be measured using these variables.
* There are different levels of measurement in statistics and data measured using them can be broadly classified into qualitative and quantitative data.
* The level of measurement of a variable decides the statistical test type to be used. The mathematical nature of a variable or in other words, how a variable is measured is considered as the level of measurement.

## What are Nominal, Ordinal, Interval and Ratio Scales?

* Nominal, Ordinal, Interval, and Ratio are defined as the four fundamental levels of measurement scales that are used to capture data in the form of [surveys](https://www.questionpro.com/blog/surveys/) and [questionnaires](https://www.questionpro.com/blog/what-is-a-questionnaire/), each being a [multiple choice question](https://www.questionpro.com/article/multiple-choice-questions.html).
* Each scale is an incremental level of measurement, meaning, each scale fulfils the function of the previous scale



### a) Nominal /Categorical Scale: 1st Level of Measurement

* [Nominal Scale,](https://www.questionpro.com/blog/nominal-scale/) also called the categorical variable scale, is defined as a scale used for labelling variables into distinct classifications and doesn’t involve a quantitative value or order.

For a question such as:

|  |  |  |
| --- | --- | --- |
| What is your Gender? | What is your Political preference? | Where do you live? |
| * M- Male * F- Female | * 1- Independent * 2- Democrat * 3- Republican | * 1- Suburbs * 2- City * 3- Town |

Nominal scale is often used in research surveys and questionnaires where only variable labels hold significance.

For instance, a customer survey asking “Which brand of smartphones do you prefer?”

Options :

“Apple”- 1

“Samsung”-2

“OnePlus”-3

* In this [survey question](https://www.questionpro.com/article/survey-question-answer-type.html), only the names of the brands are significant for the researcher conducting consumer research.
* There is no need for any specific order for these brands.
* However, while capturing nominal data, researchers conduct analysis based on the associated labels.
* In the above example, when a survey respondent selects Apple as their preferred brand, the data entered and associated will be “1”.
* This helped in quantifying and answering the final question – How many respondents selected Apple, how many selected Samsung, and how many went for OnePlus – and which one is the highest.
* This is the fundamental of quantitative research, and nominal scale is the most fundamental research scale.

### b) Ordinal Scale: 2nd Level of Measurement

* [Ordinal Scale](https://www.questionpro.com/blog/ordinal-scale/) is defined as a variable measurement scale used to simply depict the order of variables and not the difference between each of the variables.
* These scales are generally used to depict non-mathematical ideas such as frequency, satisfaction, happiness, a degree of pain, etc.

For example, a semantic differential scale question such as:

How satisfied are you with our services?

* Very Unsatisfied – 1
* Unsatisfied – 2
* Neutral – 3
* Satisfied – 4
* Very Satisfied – 5
* Here, the order of variables is of prime importance and so is the labelling. Very unsatisfied will always be worse than unsatisfied and satisfied will be worse than very satisfied.
* This is where ordinal scale is a step above nominal scale – the order is relevant to the results and so is their naming.

### c) Interval Scale: 3rd Level of Measurement

* In **interval** measurement the distance between attributes does have meaning.
* For example, when we measure temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80.
* An interval scale is one where there is order and the difference between two values is meaningful.
* You can categorize, rank, and infer equal intervals between neighboring data points, but there is no true zero point.

The following questions fall under the Interval Scale category:

* What is your family income?
* What is the temperature in your city?

### Ratio Scale: 4th Level of Measurement

* [Ratio Scale](https://www.questionpro.com/blog/ratio-scale/) is defined as a variable measurement scale that not only produces the order of variables but also makes the difference between variables known along with information on the value of true zero.
* It is calculated by assuming that the variables have an option for zero, the difference between the two variables is the same and there is a specific order between the options.

#### Ratio Scale Examples

The following questions fall under the Ratio Scale category:

* What is your daughter’s current height?
  + Less than 5 feet.
  + 5 feet 1 inch – 5 feet 5 inches
  + 5 feet 6 inches- 6 feet
  + More than 6 feet
* What is your weight in kilograms?
  + Less than 50 kilograms
  + 51- 70 kilograms
  + 71- 90 kilograms
  + 91-110 kilograms
  + More than 110 kilograms

### Summary –  Levels of Measurement

The four data measurement scales – nominal, ordinal, interval, and ratio –  are quite often discussed in academic teaching. Below easy-to-remember chart might help you in your statistics test.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Offers:** | **Nominal** | **Ordinal** | **Interval** | **Ratio** |
| The sequence of variables is established | – | Yes | Yes | Yes |
| Median | – | Yes | Yes | Yes |
| Mean | – | – | Yes | Yes |
| Difference between variables can be evaluated | – | – | Yes | Yes |
| Addition and Subtraction of variables | – | – | Yes | Yes |
| Absolute zero | – | – | – | Yes |

**PROBABILITY**

# **Basic Concepts of Probability**

* A probability is a number that reflects the chance or likelihood that a particular event will occur.
* Probabilities can be expressed as proportions that range from 0 to 1, and they can also be expressed as percentages ranging from 0% to 100%.
* A probability of 0 indicates that there is no chance that a particular event will occur, whereas a probability of 1 indicates that an event is certain to occur.
* A probability of 0.45 (45%) indicates that there are 45 chances out of 100 of the events occurring.

The concept of probability can be illustrated in the context of a study of obesity in children 5-10 years of age who are seeking medical care at a particular paediatric practice. The population (sampling frame) includes all children who were seen in the practice in the past 12 months and is summarized below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Age (years)** | | | | | |  |
|  | **5** | **6** | **7** | **8** | **9** | **10** | **Total** |
| **Boys** | 432 | 379 | 501 | 410 | 420 | 418 | 2,560 |
| **Girls** | 408 | 513 | 412 | 436 | 461 | 500 | 2,730 |
| **Totals** | 840 | 892 | 913 | 846 | 881 | 918 | 5,290 |

## Unconditional Probability

* If we select a child at random (by simple random sampling), then each child has the same probability (equal chance) of being selected, and the probability is 1/N, where N=the population size.
* Thus, the probability that any child is selected is 1/5,290 = 0.0002.
* In most sampling situations we are generally not concerned with sampling a specific individual but instead we concern ourselves with the probability of sampling certain types of individuals.
* For example, what is the probability of selecting a boy or a child 7 years of age?
* The following formula can be used to compute probabilities of selecting individuals with specific attributes or characteristics.

**P(characteristic) = # persons with characteristic / N**

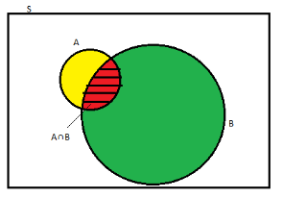
Try to figure these out before looking at the answers:

1. What is the probability of selecting a boy?
2. What is the probability of selecting a 7 year-old?
3. What is the probability of selecting a boy who is 10 years of age?
4. What is the probability of selecting a child (boy or girl) who is at least 8 years of age?
5. If we select a child at random, the probability that we select a boy is computed as follows P(boy) = 2,560/5,290 = 0.484 or 48.4%.
6. The probability of selecting a child who is 7 years of age is P(7 years of age) = 913/5,290 = 0.173.
7. P(boy who is 10 years of age) = 418/5,290 = 0.079.
8. P(at least 8 years of age) = (846 + 881+ 918)/5,290 = 2,645/5,290 = 0.500.

# **Conditional Probability**

**Conditional probability** is known as the possibility of an event or outcome happening, based on the existence of a previous event or outcome. It is calculated by multiplying the probability of the preceding event by the renewed probability of the succeeding, or conditional, event.

The probability of occurrence of any event A when another event B in relation to A has already occurred is known as conditional probability. It is depicted by P(A|B).



As depicted by the above diagram, sample space is given by S, and there are two events A and B. In a situation where event B has already occurred, then our sample space S naturally gets reduced to B because now the chances of occurrence of an event will lie inside B.

As we have to figure out the chances of occurrence of event A, only a portion common to both A and B is enough to represent the probability of occurrence of A, when B has already occurred. The common portion of the events is depicted by the intersection of both the events A and B, i.e. A ∩ B.

## Formula

When the intersection of two events happen, then the [formula for conditional probability](https://byjus.com/conditional-probability-formula/) for the occurrence of two events is given by;

**P(A|B) =**

**Or**

**P(B|A) = N(A∩B)/N(A)**

Where P(A|B) represents the probability of occurrence of A given B has occurred.

N(A ∩ B) is the number of elements common to both A and B.

N(B) is the number of elements in B, and it cannot be equal to zero.

**Example 1: Two dies are thrown simultaneously, and the sum of the numbers obtained is found to be 7. What is the probability that the number 3 has appeared at least once?**

Solution: The sample space S would consist of all the numbers possible by the combination of two dies. Therefore S consists of 6 × 6, i.e. 36 events.

Event A indicates the combination in which 3 has appeared at least once.

Event B indicates the combination of the numbers which sum up to 7.

A = {(3, 1), (3, 2), (3, 3)(3, 4)(3, 5)(3, 6)(1, 3)(2, 3)(4, 3)(5, 3)(6, 3)}

B = {(1, 6)(2, 5)(3, 4)(4, 3)(5, 2)(6, 1)}

P(A) = 11/36

P(B) = 6/36

A ∩ B = 2

P(A ∩ B) = 2/36

Applying the conditional probability formula we get,

P(A|B) = P(A∩B)/P(B) = (2/36)/(6/36) = ⅓

**Example 2: The table below shows the occurrence of diabetes in 100 people. Let D and N be the events where a randomly selected person "has diabetes" and "not overweight". Then find P(D | N).**

|  |  |  |
| --- | --- | --- |
|  | Diabetes (D) | No Diabetes (D') |
| Not overweight (N) | 5 | 45 |
| Overweight (N') | 17 | 33 |

**Solution:**

From the given table, P(N) = (5+45) / 100 = 50/100.

P(D ∩ N) = 5/100.

By the conditional probability formula,

P(D | N) = P(D ∩ N) / P(N)

= (5/100) / (50/100)

= 5/50

= 1/10

**Answer: P(D | N) = 1/10.**

**Example 3: The probability that it will be sunny on Friday is 4/5. The probability that an ice cream shop will sell ice creams on a sunny Friday is 2/3. Then find the probability that it will be sunny and the ice cream shop sells the ice creams on Friday.**

**Solution:**

Let us assume that the probabilities for a Friday to be sunny and for the ice cream shop to sell ice creams be S and I respectively. Then,

P(S) = 4/5.

P(I | S) = 2/3.

We have to find P(S ∩ I).

We can see that S and I are dependent events. By using the dependent events' formula of conditional probability,

P(S ∩ I) = P(I | S) · P(S) = (2/3) · (4/5) = 8/15.

**Answer: The required probability = 8/15.**

**Bayes' Theorem**

Bayes' Theorem is a way of finding a [probability](https://www.mathsisfun.com/data/probability.html) when we know certain other probabilities.

The formula is:

P(A|B) =

|  |  |  |
| --- | --- | --- |
| Which tells us: |  | how often A happens *given that B happens*, written **P(A|B)**, |
| When we know: |  | how often B happens *given that A happens*, written **P(B|A)** |
|  |  | and how likely A is on its own, written **P(A)** |
|  |  | and how likely B is on its own, written **P(B)** |

**Example:**

Let us say P(Fire) means how often there is fire, and P(Smoke) means how often we see smoke, then:

P(Fire|Smoke) means how often there is fire when we can see smoke  
P(Smoke|Fire) means how often we can see smoke when there is fire

So the formula kind of tells us "forwards" P(Fire|Smoke) when we know "backwards" P(Smoke|Fire)

### Example:

### dangerous fires are rare (1%)

### but smoke is fairly common (10%) due to barbecues,

### and 90% of dangerous fires make smoke

We can then discover the **probability of dangerous Fire when there is Smoke**:

P(Fire|Smoke) =P(Fire) P(Smoke|Fire)/**P(Smoke)**

=1% x 90% / **10% = 0.01\*0.9 /0.1 = 0.09 = 9%**

=9%

So it is still worth checking out any smoke to be sure.

### Example: Picnic Day

You are planning a picnic today, but the morning is cloudy

* Oh no! 50% of all rainy days start off cloudy!
* But cloudy mornings are common (about 40% of days start cloudy)
* And this is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)

**What is the chance of rain during the day?**

We will use Rain to mean rain during the day, and Cloud to mean cloudy morning.

The chance of Rain given Cloud is written P(Rain|Cloud)

So let's put that in the formula:

P(Rain|Cloud) = P(Rain) P(Cloud|Rain)/ **P(Cloud)**

* P(Rain) is Probability of Rain = 10%
* P(Cloud|Rain) is Probability of Cloud, given that Rain happens = 50%
* P(Cloud) is Probability of Cloud = 40%

P(Rain|Cloud) = 0.1 x 0.5 / **0.4**  = .125

Or a 12.5% chance of rain. Not too bad, let's have a picnic!

# **Random Variable**

Sashi plans to begin selling ice-creams in the month of March. She manufactures a batch of 1200 waffle-cones for the same. She then examines random waffle-cones and judges each cone as either “defective” or “non-defective”. She decides to examine 3 waffle-cones.

* The process of observation of activity is termed as an **experiment**.
* The results of the observation are termed as **outcomes** of the experiment.
* **Random experiments** are those experiments whose outcomes can't be predicted.

Examining a random waffle-cone and judging each cone as either defective or non-defective, is the experiment in the above scenario.

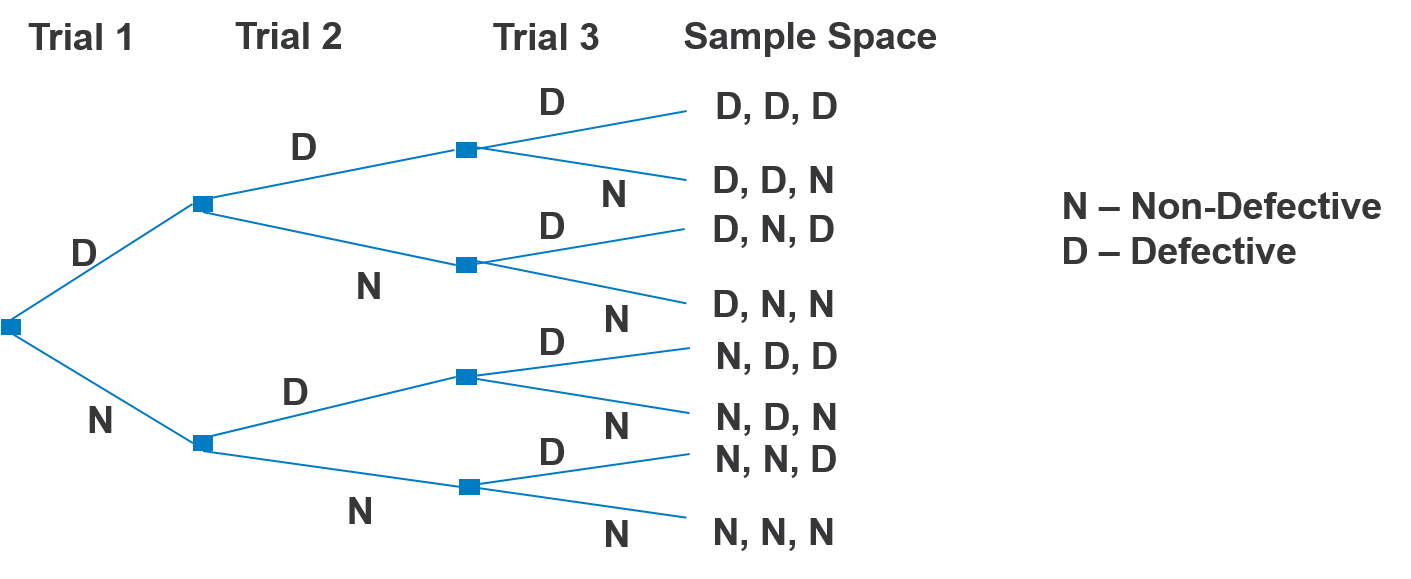
The outcome of the above experiment can be either defective(**D**) or non-defective(**N**).

Since the exact outcome of the experiment (**D** or **N**) cannot be predicted, the above experiment is a random experiment.

* Individual repetitions of the same random experiment are termed as the**trial**.
* The set of all possible outcomes in a random experiment is called **Sample space**.

Examining a waffle-cone is a trial within the experiment composed of examining 3 waffle cones.

The sample space can be defined for the experiment using a tree diagram as shown below:



One particular outcome or a set of some outcomes from the entire sample space is termed as an **event**.

In the experiment, the event of interest can be composed of all the outcomes in which the total no. of defective cones is two. Then the event **E**is

**E = {DDN, DND, NDD}**

Hence, there are 3 ways in which the above event can occur.

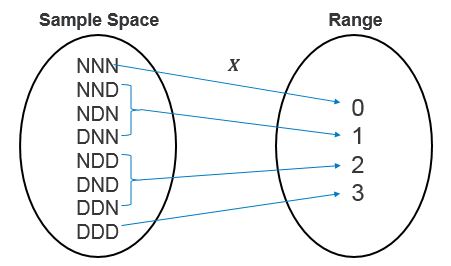
In the experiment, the event of interest can be composed of all the outcomes in which the total no. of defective cones is one or more. Then the event**E**is

**E = {DDD, DNN, NDN, NND, DDN, DND, NDD}**

Hence, there are 7 ways in which the above event can occur.

A function **X** can be defined or explained on the Sample Space as a relation where each independent sample space outcome is mapped to a numerical value, based on the event of interest.

For example, if you go through the total number of defective cones in 3 trials then, **X** can be depicted as follows:



**X** is called as **Random Variable**.

* A real-valued function, defined over a sample space is called a random variable.
* Only one real value is assigned by function to each individual outcome.
* **X** or any other uppercase letter denotes a random variable.
* A lowercase letter is used, for example, **x**  denotes the real value that can be mapped with a random variable map and its each outcome inside the sample space.
* **X** is not a variable like in Algebra. e.g. x+2 = 7, where x value is unknown.
* **X** is a function and can be depicted as follows:

**X = {x1, x2, x3, ....} or X = xk, where k = 1,2,3,...**

You can explain random variable **X** in the ice-cream scenario as follows:

X = Total no. of defective cones in 3 trials = {0,1,2,3}

You may know/be aware of the possible values of X i.e. 0,1,2,3 before to begin the experiment, you cannot be sure what values will be taken at the experiment end.

Also, **X**can assume different values each time the experiment is performed.

Due to its trial to trial variability in value and non-predictable nature, **X** is called a random variable.

# **Types of Random Variables**

A random variable may either be **Discrete** or **Continuous**.

* A random variable having a finite range is called a **Discrete Random Variable**.

For example, the total number of scratches found on a glass surface or proportion or percentage of defective boxes among 1000 tested

In real time experiments, you could record the total no of transmitted bits in error transmissions as integers or in fractions like 0.0042 proportion of the total 10,000 transmitted bits. But the fractional count can be put as numbers on a number line. So, whenever, the count is limited to a finite point on the real number line, those random variables are called discrete random variables.

* Similarly, a random variable having values within real numbers interval is called a **continuous random variable**.

For example, Height of a child measured between age 2 and 6, Weight of a person logged over a span of 2 years.

Sometimes you can observe that calculations (like heights, weights, current in a wire) assume the value in a range of the real numbers (always true theoretically). There is the possibility for an arbitrary precision in the calculation. However, in real life application, you may round off the value to a nearest 10th or 100th of a unit. Continuous random variables are such random variables that represent this type of range of values.

Sometimes, when the range of possible values is very high, you can consider a discrete random variable **X** is continuous. For example, a multi-meter output that displays the current at the nearest 100th of a milliampere. Although the measurements are limited, and it could be perceived to be discrete random variable it can also be considered as continuous random variables if needed for simplification in computation.

Possible outcomes for **X**= Number of defective cones in 3 trials

i.e., X = {0,1,2,3}

**X** is a discrete random variable because it results in values that are finite in nature and are from the above set.

In the ice-cream scenario, some of the other examples that can be considered as discrete random variables are:

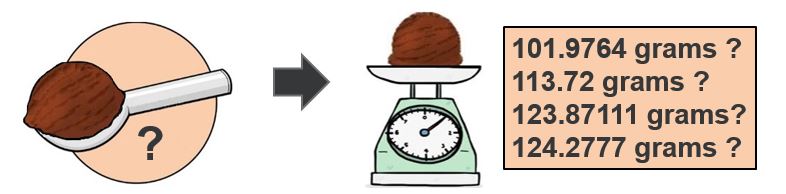
* Number of customers expected to arrive between 2 pm and 3 pm
* Number of ice-creams bought by children below 12 years of age
* Number of ice-cream cones that break/leak while being filled with ice-cream
* Number of ice-creams sold in a month
* Number of ice-creams sold in a year

Sashi uses a 4oz dish spoon to serve ice-cream equivalent to 113 grams, per scoop.

Let us consider a random variable, **X**= Weight of one scoop of ice-cream.

Let us consider, the permissible error margin is 10% for the weight of a single scoop of ice-cream, so our **X**can assume any values in the interval **[101.7 grams, 124.3 grams]**and outcomes could be

X = {101.7, 101.88, ..................., 110.346, .................. 124.29}



If you consider any two values in the interval [101.7,124.3] say, 102 and 103, here you can get any value in the interval [102, 103]. You can assume any value between 102.5 and 102.9 for the weight of a single scoop.

In such a scenario, wherein an interval, **X**can assume any value, is defined to be a continuous random variable.

Few other examples for continuous random variables :

* The exact time taken by Sashi to serve a customer in the [10 sec, 15 sec] interval with an accuracy of 100th of a second.
* The precise diameter of a single ice-cream scoop that Sashi serves, assuming the average diameter of a single scoop that Sashi serves lies in the interval [2 inches, 2.25 inches]
* The exact thickness of the waffle-sheet with which the waffle-cones are made from, assuming the average thickness of a waffle-sheet lies in the interval [1.4 mm, 1.8 mm]

# **Probability Distribution**

By studying Distributions, you can understand how the random variable behaves. When the possibility of random variable values is associated with each of its probabilities, you get its Probability Distribution.

The probability distribution is usually represented through either a table or a Graph(usually a histogram).

Recall, for a Finite Sample Space **S,**then the probability **P(A),**is a real number assigned to the event **A** such that **0<=P(A)<= 1** and **P(S)=1**

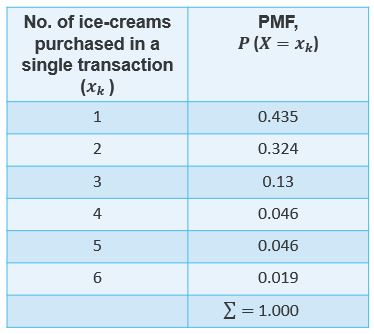
# **Types of Probability Distributions**

Probability Distributions can be Discrete or Continuous.

* The associated probability distribution for a random variable with discrete values is called a **Discrete Probability Distribution**
  + Discrete Probability Distributions are described by using the**Probability Mass Function (PMF)**.
* The associated probability distribution for a random variable with continuous (or approx. continuous) values is called a **Continuous Probability Distribution**
  + Continuous Probability Distributions are described by using the **Probability Density Function (PDF)**.

# **Probability Mass Function PMF**

In the table given below:



The probability of a customer buying exactly one ice-cream is 0.435 i.e.,**P(X=1)=0.435**i.e.,**43.5%.**

Find the probability of next customer buying exactly 4 ice-creams?

# **Types of Discrete Distributions**

* The Discrete Uniform Distribution
* The Binomial Distribution
* The Negative Binomial Distribution
* The Poisson Distribution

**Exploratory Data Analysis**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before getting them dirty with it.

The goal of EDA is to allow data scientists to get deep insight into a data set and at the same time provide specific outcomes that a data scientist would want to extract from the data set. It includes:

* List of outliers
* Estimates for parameters
* Uncertainties about those estimates
* List of all important factors
* Conclusions or assumptions as to whether certain individual factors are statistically essential
* Optimal settings
* A good predictive model

The purpose of Exploratory Data Analysis is essential to tackle specific tasks such as:

* Spotting missing and erroneous data;
* Mapping and understanding the underlying structure of your data;
* Identifying the most important variables in your dataset;
* Testing a hypothesis or checking assumptions related to a specific model;
* Establishing a model that can explain your data using minimum variables
* Estimating parameters and figuring the margins of error.

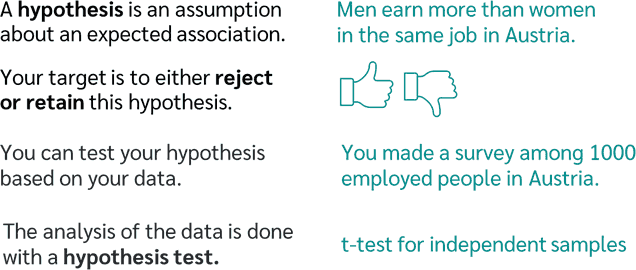
Importance of using EDA for analyzing data sets is:

* Helps identify errors in data sets.
* Gives a better understanding of the data set.
* Helps detect outliers or anomalous events.
* Helps understand data set variables and the relationship among them.
* Improve understanding of variables by extracting averages, mean, minimum, and maximum values, etc.

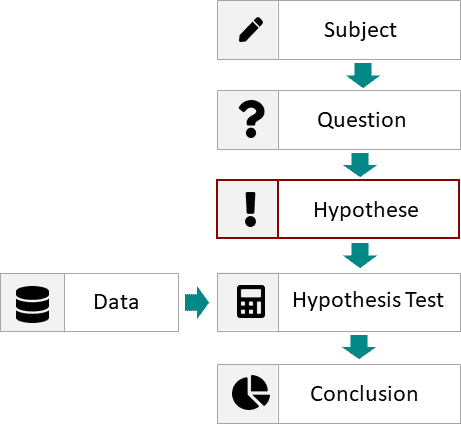
**Univariate analysis tests**

**Hypothesis Testing**

* A hypothesis is an assumption that is neither proven nor disproven.
* In the research process, a hypothesis is made at the very beginning and the goal is to either reject or not reject the hypothesis.
* In order to reject or not reject a hypothesis, data is needed, which is then evaluated using a hypothesis test.
* An example of a hypothesis could be: "Men earn more than women in the same job."



* In order to formulate a hypothesis, a research question must first be defined.
* A precisely formulated hypothesis about the population can then be derived from the research question, e.g. men earn more than women in the same job



### Null and Alternative hypothesis

There are always two hypotheses that are exactly opposite to each other, or that claim the opposite. These opposite hypotheses are called null and alternative hypothesis and are abbreviated with H0 and H1.

### Null hypothesis H0:

The null hypothesis assumes that there is no difference between two or more groups with respect to a characteristic.

###### **Example:**

The salary of men and women does not differ.

### Alternative hypothesis H1:

Alternative hypotheses, on the other hand, assume that there is a difference between two or more groups.

###### **Example:**

The salary of men and women differs.

**Level of significance**

* A hypothesis test can never reject the null hypothesis with absolute certainty.
* There is always a certain probability of error that the null hypothesis is rejected even though it is actually true.
* This probability of error is called the significance level or α.
* The significance level is used to decide whether the null hypothesis should be rejected or not.
* If the p-value is smaller than the significance level, the null hypothesis is to be rejected; otherwise, it is not to be rejected.
* Usually, a significance level of 5% or 1% is set. If a significance level of 5% is set, it means that it is 5% likely to reject the null hypothesis even though it is actually true.

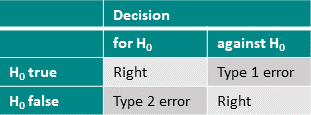
**Types of errors**

Because a hypothesis can only be rejected with a certain probability, different types of errors occur.

There are two types of errors in hypothesis testing:

**Type 1 error**: If the alternative hypothesis is accepted although the null hypothesis is valid.

**Type 2 error**: If the null hypothesis is retained although the alternative hypothesis applies.



**Multivariate Analysis**

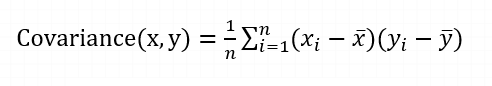
**Finding relationship in data : Covariance and Correlation**

Covariance and correlation are two terms that are opposed and are both used in [statistics](https://www.simplilearn.com/what-is-statistical-analysis-article) and [regression analysis.](https://www.simplilearn.com/tutorials/excel-tutorial/regression-analysis) Covariance shows you how the two variables differ, whereas correlation shows you how the two variables are related.

**Covariance**

* Covariance is a statistical term that refers to a systematic relationship between two random variables.
* It signifies the direction of the linear relationship between the two variables.
* By direction we mean if the variables are directly proportional or inversely proportional to each other
* The covariance value can range from -∞ to +∞, with a negative value indicating a negative relationship and a positive value indicating a positive relationship.
* The greater this number, the more reliant the relationship.
* Positive covariance denotes a direct relationship and is represented by a positive number.
* A negative number, on the other hand, denotes negative covariance, which indicates an inverse relationship between the two variables.
* Covariance is great for defining the type of relationship, but not helpful for interpreting the magnitude.

Let Σ(X) and Σ(Y) be the expected values of the variables, the covariance formula can be represented as:



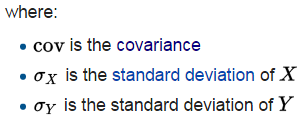
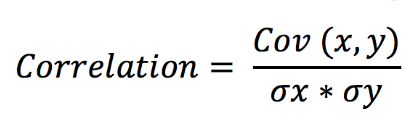
Where,

* xi = data value of x
* yi = data value of y
* x̄ = mean of x
* ȳ = mean of y
* N = number of data values.

## ****Correlation****

* Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables.
* It not only shows the kind of relation (in terms of direction) but also how strong the relationship is.
* Thus, we can say the correlation values have standardized notions, whereas the covariance values are not standardized and cannot be used to compare how strong or weak the relationship is because the magnitude has no direct significance.
* It can assume values from -1 to +1.

The main result of a correlation is called the correlation coefficient.



The correlation coefficient is a dimensionless metric and its value ranges from -1 to +1.

The closer it is to +1 or -1, the more closely the two variables are related.