**Statistics**

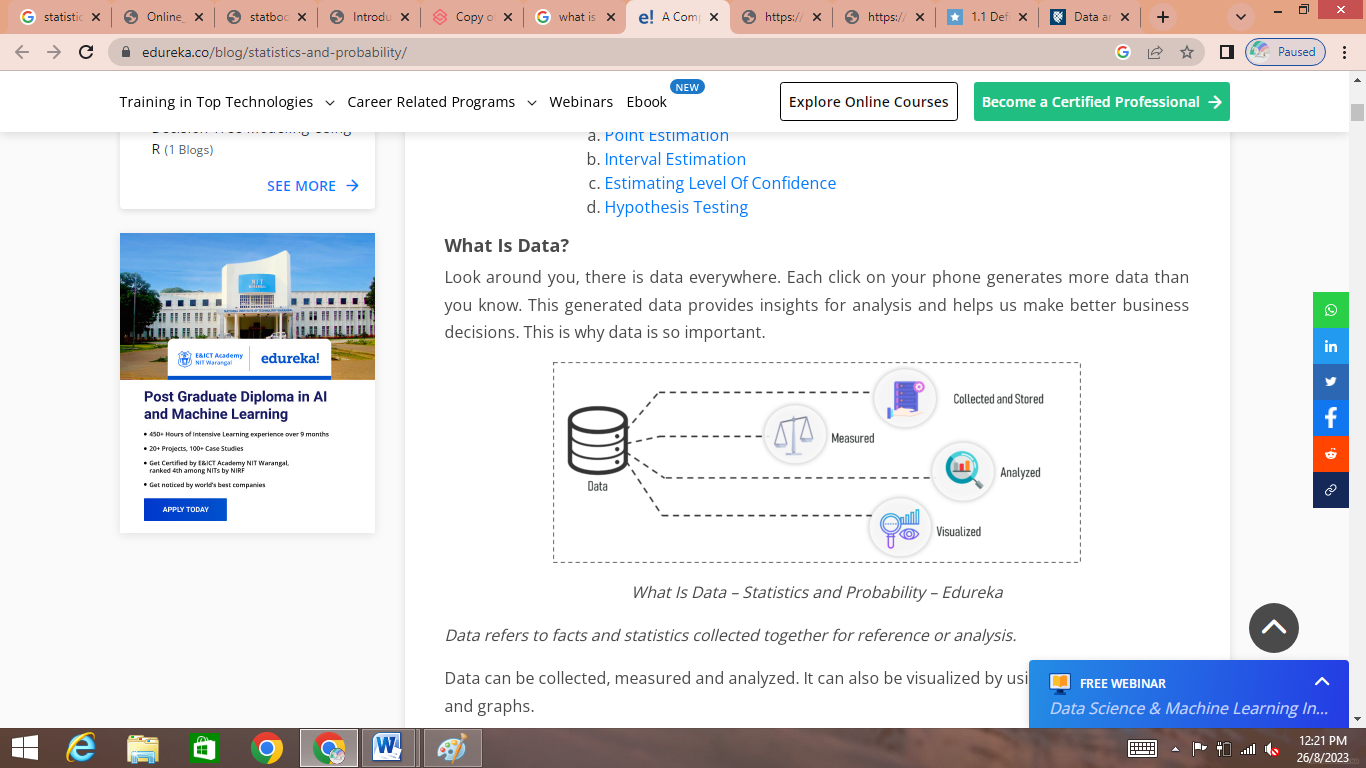
**Introduction**

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

**What Is Data?**

Look around you, there is data everywhere. Each click on your phone generates more data than you know. This generated data provides insights for analysis and helps us make better business decisions. This is why data is so important.

Data are individual pieces of factual information recorded and used for the purpose of analysis. It is the raw information from which statistics are created.



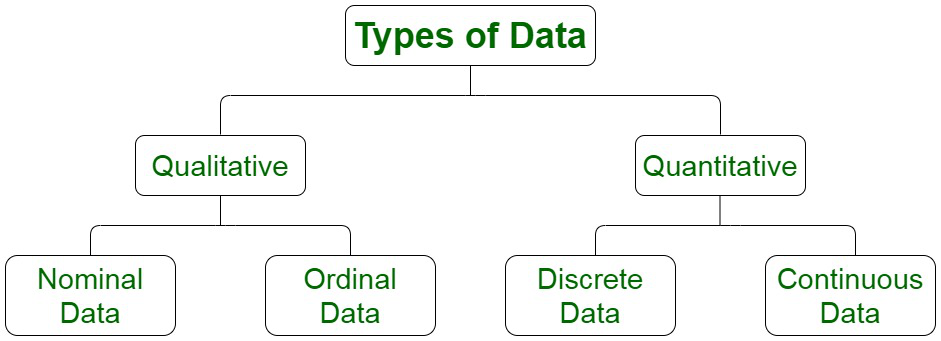
Data can be collected, measured and analyzed. It can also be visualized by using statistical models and graphs.

**Categories Of Data**

Data can be categorized into two sub-categories:

1. Qualitative Data
2. Quantitative Data

Refer the below figure to understand the different categories of data:



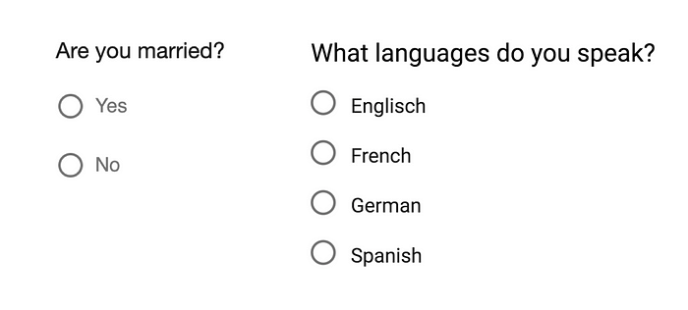
***Fig. Categories of Data***

**Qualitative Data:**

**Qualitative** data describes qualities or characteristics. It is collected using questionnaires, interviews, or observation, and frequently appears in narrative form. For example, it could be notes taken during a focus group on the quality of the food at Cafe Mac, or responses from an open-ended questionnaire.

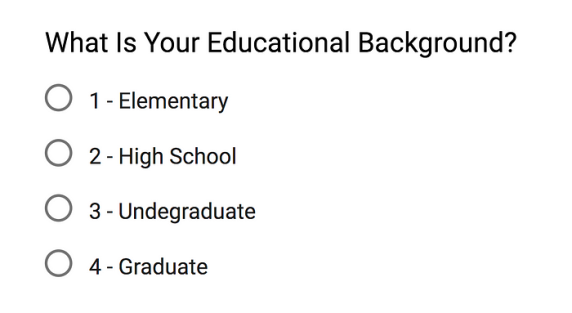
*Qualitative data is further divided into two types of data:*

* ***Nominal Data:*** Nominal data are used to label variables without any quantitative value. Common examples include male/female (albeit somewhat outdated), hair color, nationalities, names of people, and so on.



***Fig: Nominal Data***

* **Ordinal Data:** It is categorical data that can be ranked or ordered in accordance with a specific attribute or characteristic.



**Fig: Ordinal Data**

**Quantitative Data:**

Quantitative data deals with numbers and things you can measure objectively. This is further divided into two:

1. **Discrete Data:**

Also known as categorical data, it can hold a finite number of possible values. This type of data **can’t be measured but it can be counted**.

***Example:*** *Number of students in a class.*

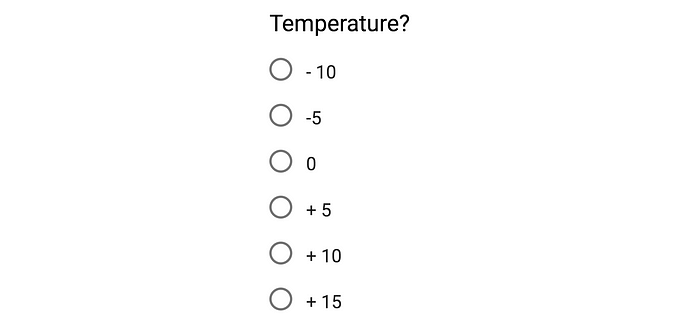
1. **Continuous Data:**

Continuous Data represents measurements and therefore their values **can’t be counted but they can be measured**. This type of data can hold an infinite number of possible values.

***Example:*** *Weight**of a person, height of a person*

1. **Interval Data**

Interval values represent **ordered units that have the same difference.** Therefore we speak of interval data when we have a variable that contains numeric values that are ordered and where we know the exact differences between the values. An example would be a feature that contains temperature of a given place like you can see below:



***Fig: Interval Data***

1. **Ratio Data**

Ratio values are also ordered units that have the same difference. Ratio values are **the same as interval values, with the difference that they do have an absolute zero**. Good examples are height, weight, length etc.

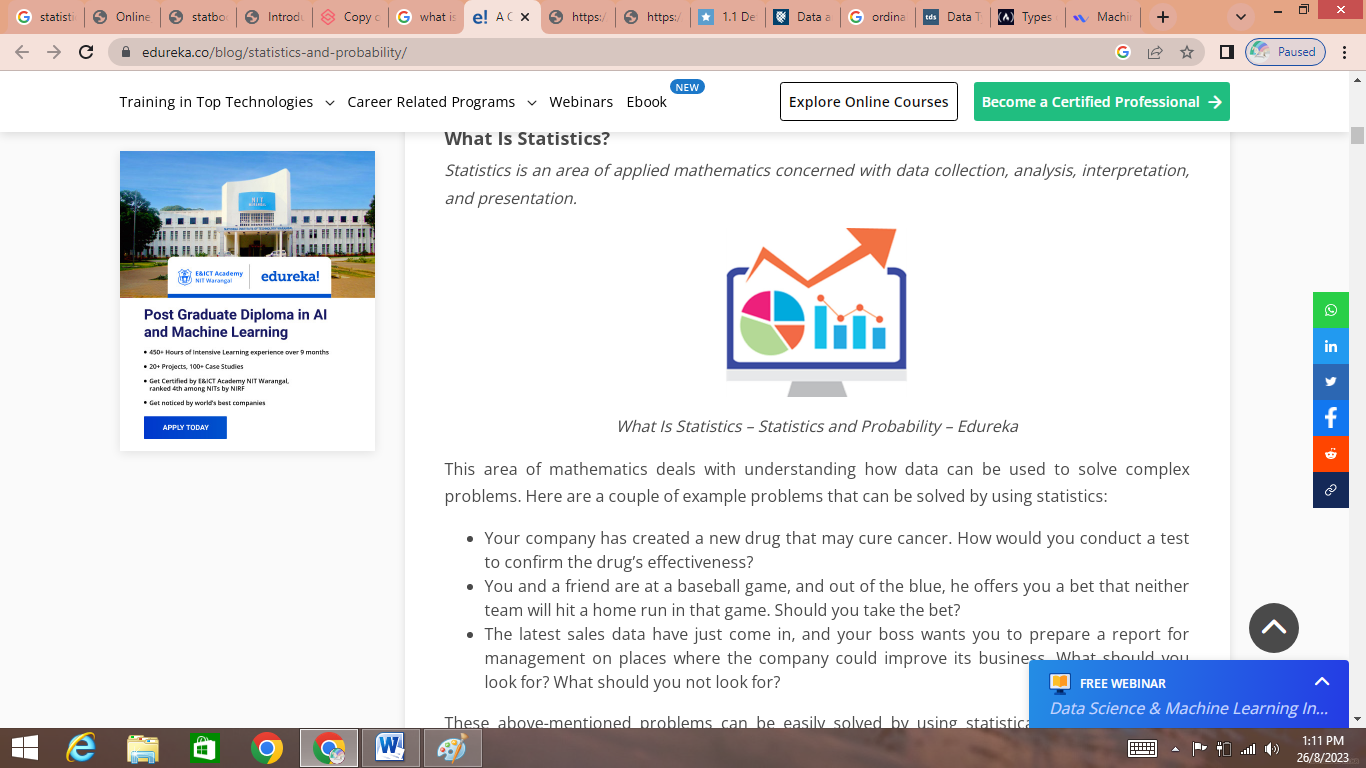


***Fig: Ratio Data***

*So these were the different categories of data.*

**Statistics:**

Statistics is a mathematical science that is concerned with the collection, analysis, interpretation or explanation, and presentation of data.



Statistics is used to process complex problems in the real world so that Data Scientists and Analysts can look for meaningful trends and changes in Data. In simple words, Statistics can be used to derive meaningful insights from data by performing mathematical computations on it.

Several Statistical functions, principles and algorithms are implemented to [analyze raw data](https://panoply.io/analytics-stack-guide/preparing-data-for-self-service-analytics/), build a Statistical Model and infer or predict the result.



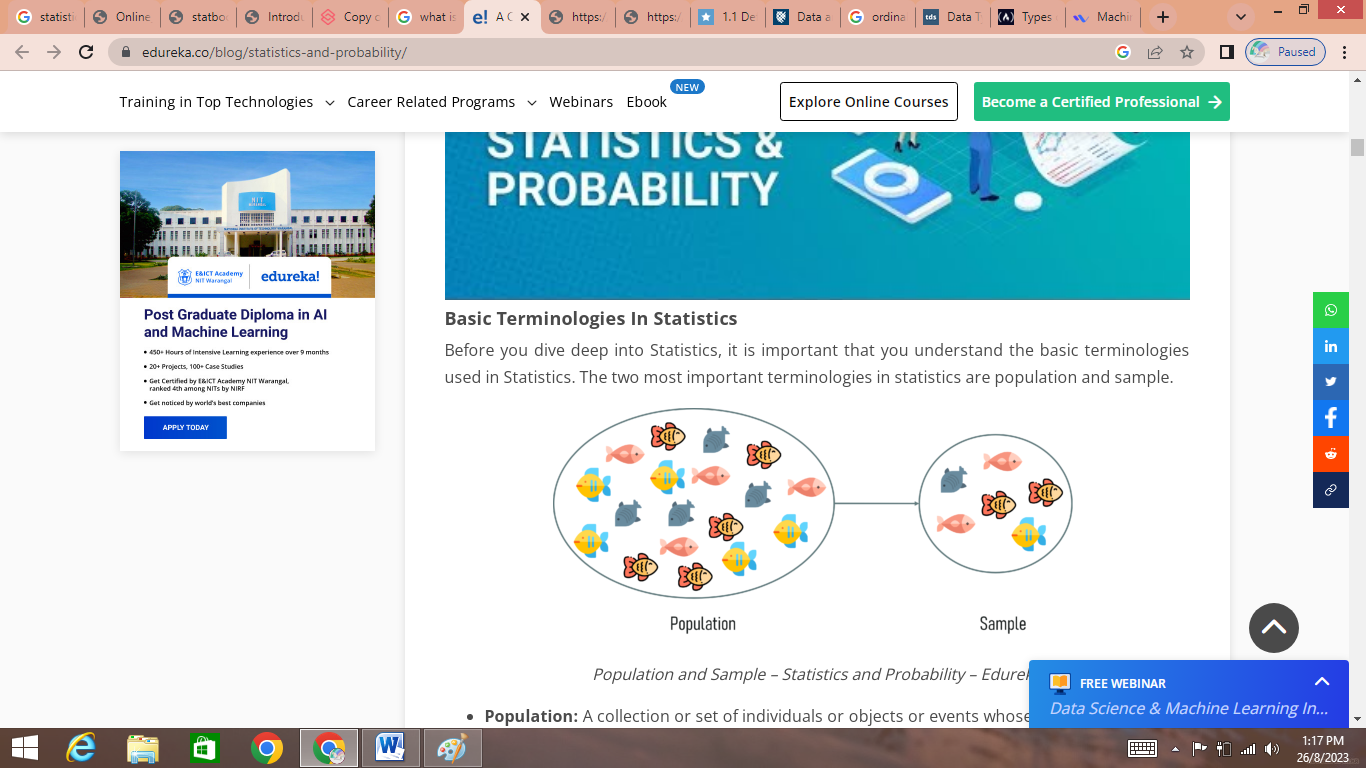
This area of mathematics deals with understanding how data can be used to solve complex problems. Here are a couple of example problems that can be solved by using statistics:

* Your company has created a new drug that may cure cancer. How would you conduct a test to confirm the drug’s effectiveness?
* You and a friend are at a baseball game, and out of the blue, he offers you a bet that neither team will hit a home run in that game. Should you take the bet?
* The latest sales data have just come in, and your boss wants you to prepare a report for management on places where the company could improve its business. What should you look for? What should you not look for?

These above-mentioned problems can be easily solved by using statistical techniques.

## ****Basic Terminologies in Statistics****

Before you dive deep into Statistics, it is important that you understand the basic terminologies used in Statistics. The two most important terminologies in statistics are population and sample.



***Fig : Population and Sample***

* **Population:** A collection or set of individuals or objects or events whose properties are to be analyzed
* **Sample:** A subset of the population is called ‘Sample’. A well-chosen sample will contain most of the information about a particular population parameter.

Now you must be wondering how can one choose a sample that best represents the entire population.

**Sampling Techniques**

*Sampling is a statistical method that deals with the selection of individual observations within a population*. It is performed to infer statistical knowledge about a population.

Consider a scenario wherein you’re asked to perform a survey about the eating habits of teenagers in the US. There are over 42 million teens in the US at present and this number is growing. Is it possible to survey each of these 42 million individuals about their health? Obviously not! That’s why sampling is used. It is a method wherein a sample of the population is studied in order to draw inference about the entire population.

There are two main types of sampling techniques:

1. Probability Sampling
2. Non-Probability Sampling

**Probability Sampling:**

This is a sampling technique in which samples from a large population are chosen using the theory of probability. In this method, every individual of the population has an equal chance of being selected. The advantage of using probability sampling is that probability sampling gives us the best chance to draw a sample that is true representative of the population. However, this method is more time consuming and expensive than the non-probability sampling method.

There are three types of probability sampling:

#### ****Random Sampling****

With random sampling, every member of the population has an equal opportunity to be included in the sample, and pure chance is the only factor that determines who actually goes into the sample.

******

***Fig: Random Sampling***

For example, we want to choose 200 students at random from a school. Here, we can give each student in the database of the school a number between 1 and 500 and choose a random sample of 200 numbers using a random number generator.

Another example is going to a shopping mall and asking every fifth person for his or her opinion about inflation.

There are two types of simple random sampling.

1. Simple Random Sampling Without Replacement(SRSWOR)
2. Simple Random Sampling With Replacement(SRSWR)

**Sample R code is below:**

**Syntax:** sample (vector\_of\_values)

**Ex1:**

>sample (c(1:10))

This request returns the following:

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

**Ex2:** Draw a simple random sample of size 5 from a population of 30 units by SRSWR and SRSWOR methods

**Solution: R- Code with output**

> swr=sample(30,5,replace=T)

> swr

[1] 1 5 2 5 25

*Sample consists of units numbered 1,5,2,5,25 by SRSWR method*

> swor=sample(30,5)

> swor

[1] 7 15 29 1 11

*Sample consists of units numbered 7,15,29,1,11 by SRSWOR method*

**Ex3.**Monthly consumption of milk (in liter) of 30 families is given below:

32,28,37,40,45,60,50,36,30,27,45,40,42,45,46,50,25,28,32,30,31,33,35,42,47,55,52,48,48,42

Draw a simple random sample of size 9 using SRSWOR method.

**Solution: R- Code with output**

>x=c(32,28,37,40,45,60,50,36,30,27,45,40,42,45,46,50,25,28,32,30,31,33,35,42,47,55,52,48,48,42)

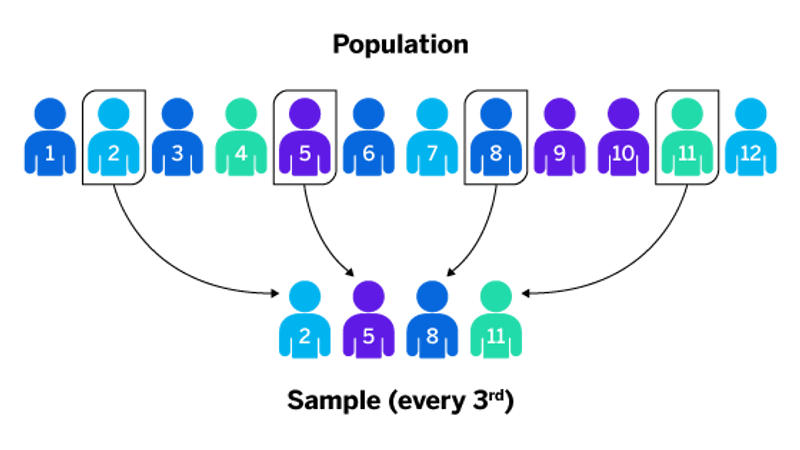
> s=sample(x,9)

> s #s contains required sample of 9 units.

[1] 37 40 45 30 31 35 42 47 28

1. ***Systematic Sampling***

In systematic sampling, the first member of sample is chosen randomly, and the others are selected according to a predetermined sampling interval (i.e., nth number). The sampling interval (i.e., nth number) is calculated by dividing the total population size by the desired sample size. Refer the below figure to better understand how Systematic sampling works.



***Fig: Systematic Sampling***

For example, if you had a list of 1,000 customers (your target population) and you wanted to survey 200 of them, your sampling interval would be one-fifth. This means that you would sample every **5th** person in your list of 1,000 customers.

**Ex1.** Draw a systematic sample of size 6 from a population of 30 units

**Solution: R- Code with output**

> ps=30

> ss=6

> k=ps/ss

> k

[1] 5

> r=sample(1:k,1)

> s=seq(r,ps,k)

> r

[1] 2

> s

[1] 2 7 12 17 22 27

# s contains unit numbers from population to be selected in sample.

9

**Ex 2:** Following data relate to number of admissions cancelled in a day for 20 days. Draw a systematic sample of size 5 from it.

7,9,13,5,4,8,11,10,9,6,5,2,8,16,5,12,10,5,8,10

**Solution: R- Code with output**

> x = c(7,9,13,5,4,8,11,10,9,6,5,2,8,16,5,12,10,5,8,10)

> ps = 20;ss=5;k=ps/ss

>k

[1] 4

> r=sample(1:k,1)

> r

[1] 4

> s=seq(r,ps,k)

>4 8 12 16 20

> rs =x[s]

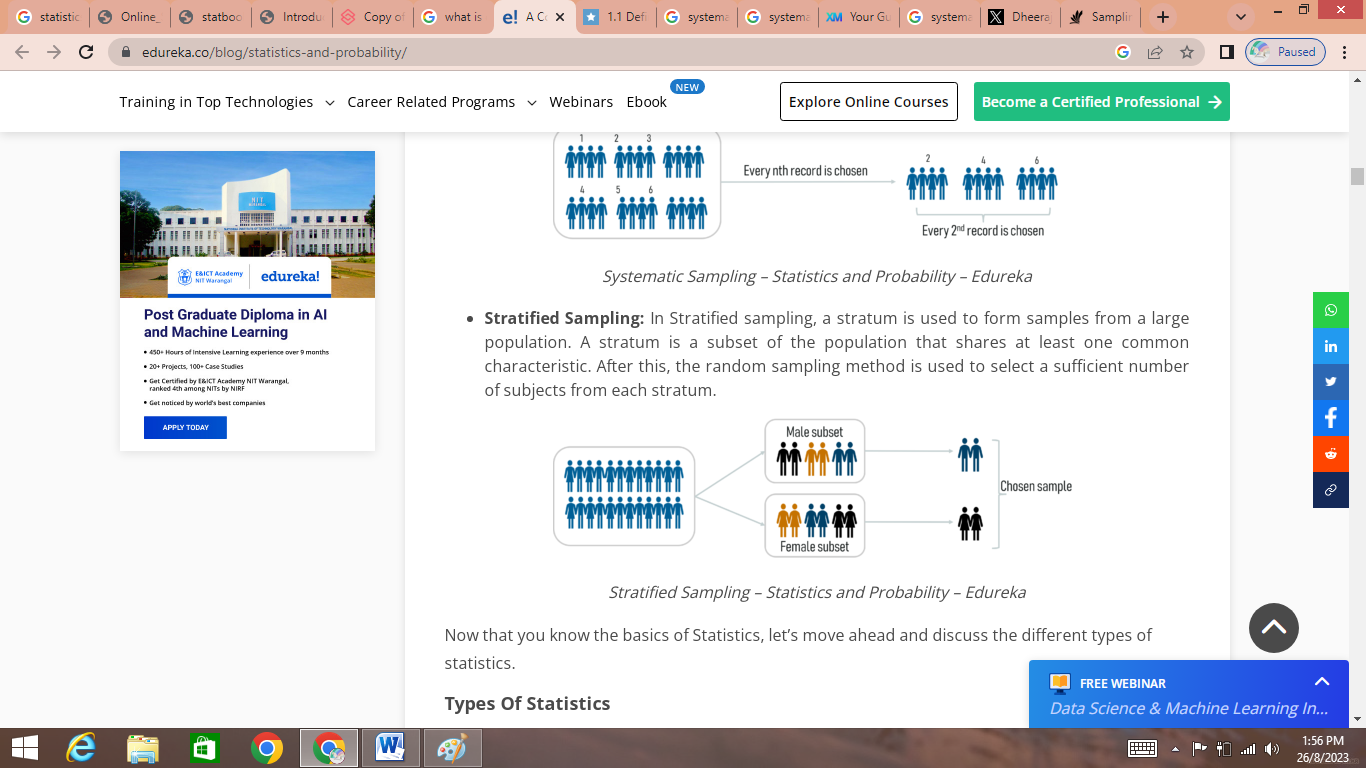
> rs

[1] 5 10 2 12 10

#rs contains required sample

1. ***Stratified Sampling***

 In Stratified sampling, a stratum is used to form samples from a large population. A stratum is a subset of the population that shares at least one common characteristic. After this, the random sampling method is used to select a sufficient number of subjects from each stratum.



***Fig: Stratified Sampling***

A good example of stratified sampling would be to divide the population into male and female or into rural and urban etc. It is useful only when you plan to subdivide the subjects for subsequent analysis to make various comparisons and decisions.

**Ex1:** Following are data on number of students enrolled in 3 years of degree course in a certain college.

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | F.Y | S.Y | T.Y |
| **No. of students** | 700 | 500 | 410 |

Draw a stratified sample of size 20, 15 and 10 from each of the class.

**Solution: R- Code with output**

> st=c(700,500,410)

> st1=sample(st[1],20)

> st1

[1] 210 472 502 123 45 57 152 79

[9] 452 197 154 619 520 654 145 152

[17]154 319 4 16

This is a sample of size 20 bearing student numbers of class F.Y.

> st2=sample(st[2],15)

> st3=sample(st[3],10)

st2,st3 vectors contains samples from S.Y, T.Y class respectively. Every time new sample will be generated.

**Ex2:** Using following data, draw a stratified sample of size 50 from different strata by method of proportional allocation

|  |  |
| --- | --- |
| **Stratum** | **Size(Ni)** |
| 1 | 400 |
| 2 | 300 |
| 3 | 150 |

**Solution: R- Code with output**

> st=c(400,300,150)

> ps=sum(st)

> ps

[1] 850

> ss=round((st/ps)\*50)

> ss

[1] 24 18 9

> st1=sample(st[1],ss[1])

> st1

[1] 142 187 263 94 79 219 213 253 265 318 320 374 33 128 223 384 252 381 30 231 108 192 86 54

> st2=sample(st[2],ss[2])

> st2

[1] 39 97 215 295 7 60 260 283 239 8 200 217 142 140 262 267 137 261

> st3=sample(st[3],ss[3])

> st3

[1] 9 5 94 73 31 12 84 30 111

>

Note that every time when sample command is issued we get difference sample

**Ex3:** Following data give monthly consumption of petrol for employees in 4 departments of a company.

|  |  |  |
| --- | --- | --- |
| **Department** | **No. of Employees** | **Monthly consumption of petrol (in liters)** |
| 1 | 5 | 10,8,12,7,9 |
| 2 | 7 | 8,7,11,20,6,8,13 |
| 3 | 4 | 12,16,15,9 |
| 4 | 6 | 10,12,14,9,19,20 |

Draw a random sample of size 3, employees from each stratum.

**Solution: R- Code with output**

|  |
| --- |
| > st1=c(10,8,12,7,9)  > st2=c(8,7,11,20,6,8,13)  > st3=c(12,16,15,9)  > st4=c(10,12,14,9,19,20)  > s1=sample(st1,3)  > s1  [1] 9 7 8  > s2=sample(st2,3)  > s2  [1] 11 6 7  > s3=sample(st3,3)  > s3  [1] 12 9 15  > s4=sample(st4,3)  > s4  [1] 12 14 10  > |

**Simulation**

The command sample can also be used for simulating some random experiments.

**Ex1:** Simulate an experiment of tossing a coin 100 times and prepare its frequency table.

**Solution: R- Code with output**

|  |
| --- |
| > x=c('H','T')*# x is a vector of outcomes of the experiment*  > r=sample(x,100,replace=T)*#r is the required sample*  > r  [1] "H" "H" "H" "T" "H" "T" "H" "H" "T" "H" "T" "H" "H" "T" "T" "T" "H" "T" "T" "H" "T" "H" "T" "T" "H" "T" "T" "T" "H" "H" "H" "H" "T" "T" "T" "H" "T" "H" "H" "T"  [41] "T" "T" "T" "H" "T" "H" "H" "H" "T" "H" "H" "T" "T" "T" "T" "H" "T" "H" "T" "T" "T" "T" "H" "H" "T" "H" "T" "H" "T" "H" "H" "H" "T" "T" "T" "H" "T" "H" "T" "T"  [81] "T" "H" "T" "H" "T" "H" "T" "H" "T" "T" "T" "H" "H" "T" "H" "T" "T" "T" "T" "T"  > fr=table(r) *#table command is for preparing a frequency table*  > fr  r  H T  44 56 |

**Ex2:** Simulate an experiment of tossing a die 50 times and prepare its frequency table.

**Solution: R- Code with output**

|  |
| --- |
| > x=1:6  > r=sample(x,50,replace=T)  > r  [1] 2 4 3 4 1 5 4 2 5 6 3 3 5 2 6 6 6 4 1 4 6 5 4 2 1 6 4 3 3 6 6 3 1 2 2 1 6 4 6 4 3 3 5 6 5 2 5 5 5 1  > fr=table(r)  > fr  r  1 2 3 4 5 6  6 7 8 9 9 11 |

**Types of Statistics**

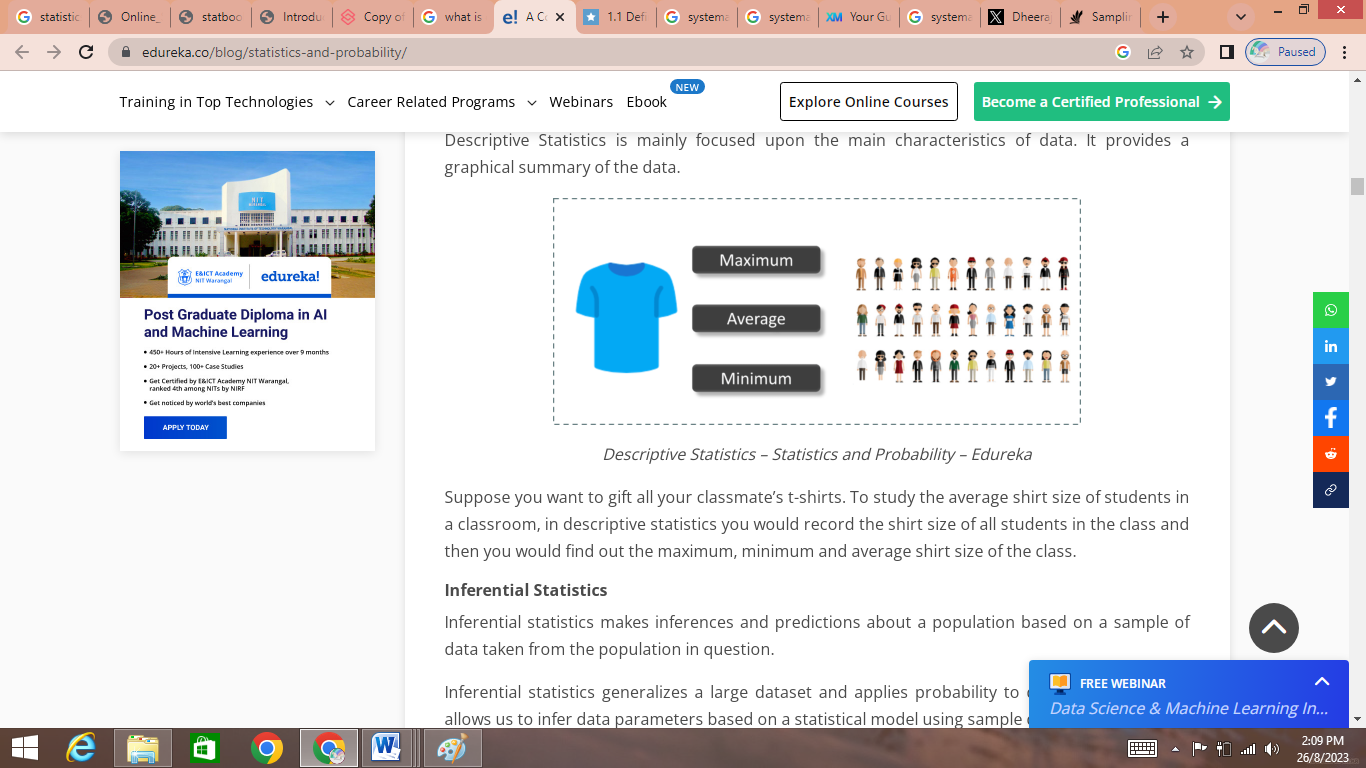
There are two well-defined types of statistics:

1. Descriptive Statistics
2. Inferential Statistics

### ****Descriptive Statistics****

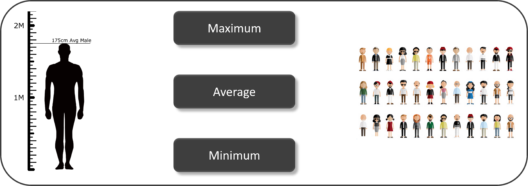
Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.

Descriptive Statistics is mainly focused upon the main characteristics of data. It provides a graphical summary of the data.



***Fig: Example of Descriptive Statistics***

Suppose you want to gift all your classmate’s t-shirts. To study the average shirt size of students in a classroom, in descriptive statistics you would record the shirt size of all students in the class and then you would find out the maximum, minimum and average shirt size of the class.



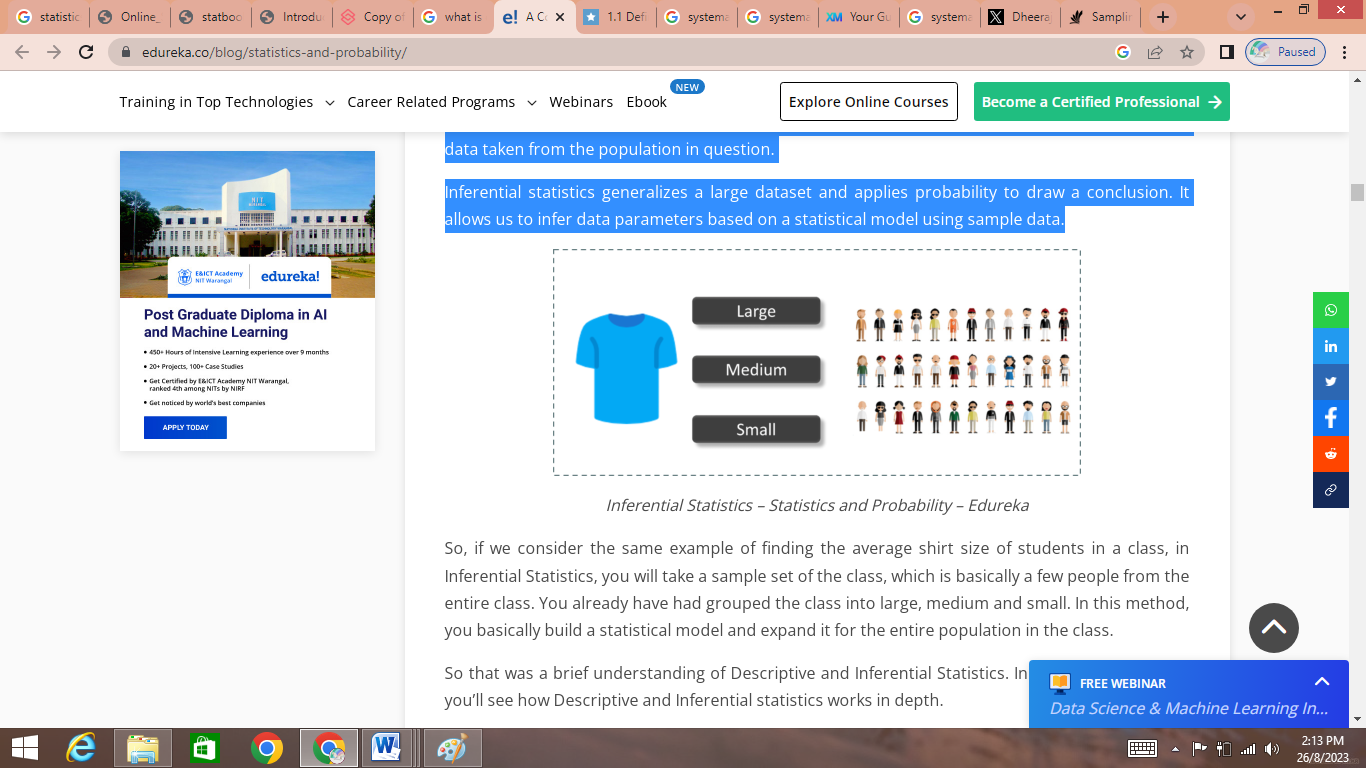
***Fig: Example of Descriptive Statistics***

Suppose you want to study the average height of students in a classroom, in descriptive statistics you would record the heights of all students in the class and then you would find out the maximum, minimum and average height of the class.

### ****Inferential Statistics****

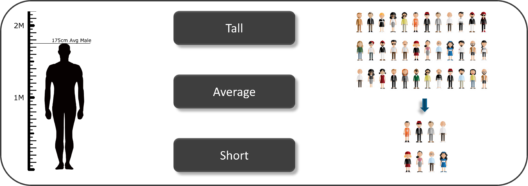
Inferential statistics makes inferences and predictions about a population based on a sample of data taken from the population in question.

Inferential statistics generalizes a large dataset and applies probability to draw a conclusion. It allows us to infer data parameters based on a statistical model using sample data.



**Fig: example of Inferential Statistics**

So, if we consider the same example of finding the average shirt size of students in a class, in Inferential Statistics, you will take a sample set of the class, which is basically a few people from the entire class. You already have had grouped the class into large, medium and small. In this method, you basically build a statistical model and expand it for the entire population in the class.



**Fig: example of Inferential Statistics**

So, if we consider the same example of finding the average height of students in a class, in Inferential Statistics, you will take a sample set of the class, which is basically a few people from the entire class. You already have had grouped the class into tall, average and short. In this method, you basically build a statistical model and expand it for the entire population in the class.

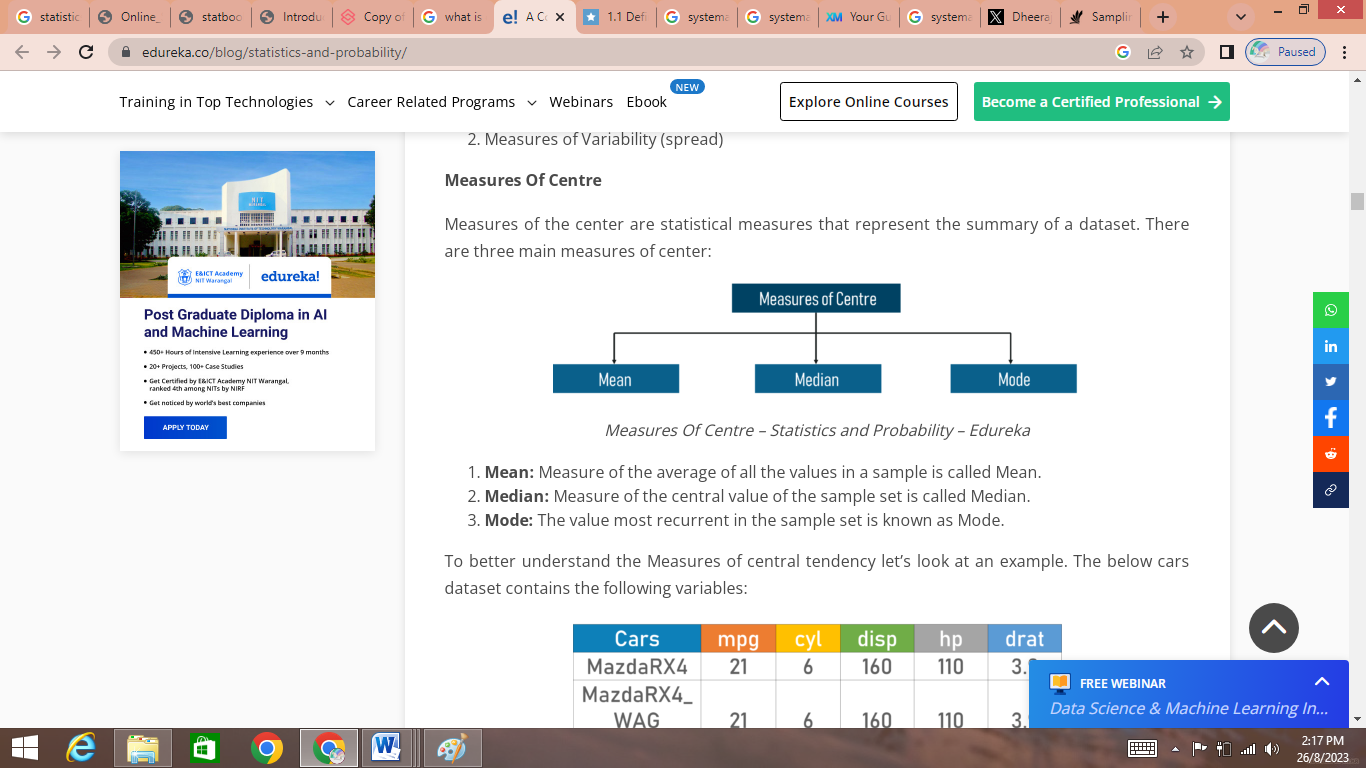
So that was a brief understanding of Descriptive and Inferential Statistics. In the further sections, you’ll see how Descriptive and Inferential statistics works in depth.

**Understanding Descriptive Statistics**

Descriptive Statistics is broken down into two categories:

1. Measures of Central Tendency
2. Measures of Variability (spread)
3. **Measures of Centre**

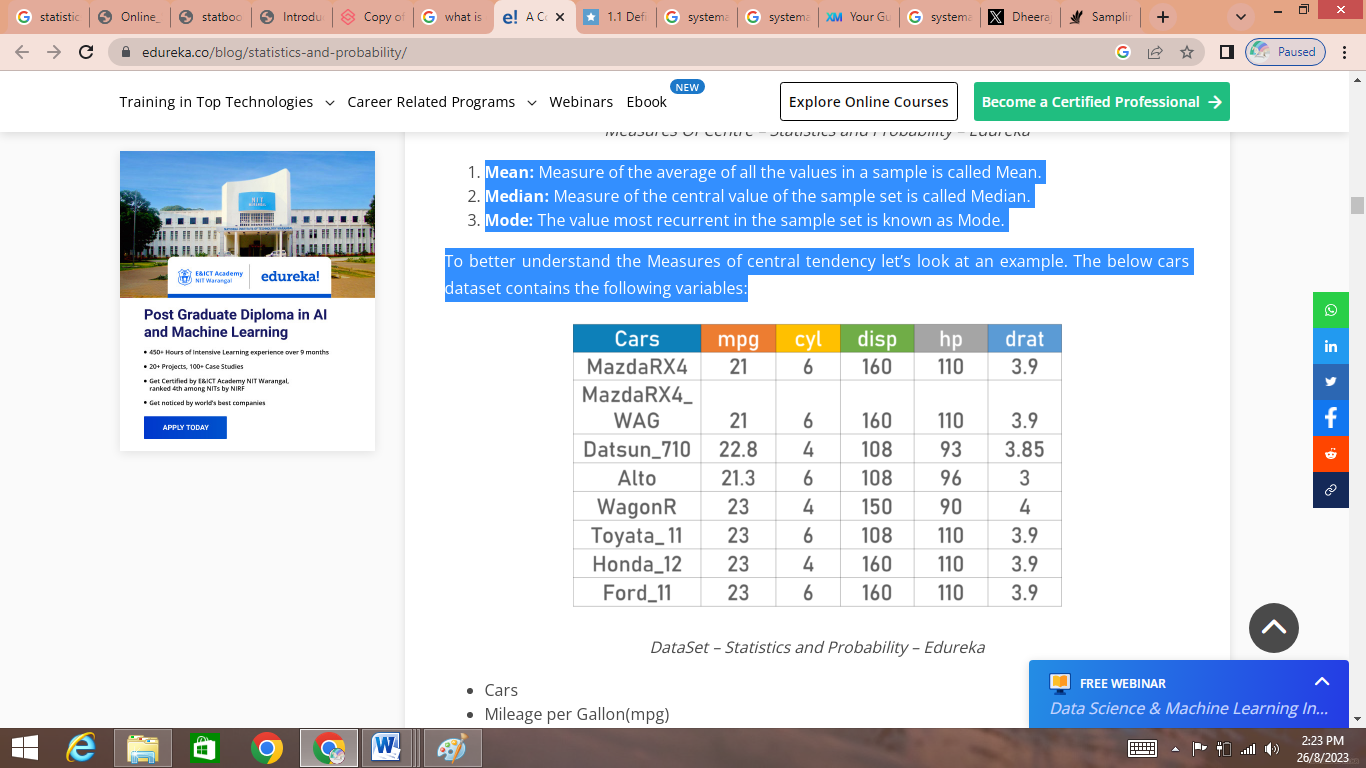
A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution. There are three main measures of center:



**Fig: Measures of Center**

1. **Mean:** The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.
2. **Median:** The median is the middle value in distribution when the values are arranged in ascending or descending order.
3. **Mode:** The mode is the most commonly occurring value in a distribution.

To better understand the Measures of central tendency let’s look at an example. The below cars dataset contains the following variables:



**Fig: Data Set**

* Cars
* Mileage per Gallon(mpg)
* Cylinder Type (cyl)
* Displacement (disp)
* Horse Power(hp)
* Real Axle Ratio(drat)

Using descriptive Analysis, you can analyze each of the variables in the sample data set for mean, standard deviation, minimum and maximum.

If we want to find out the **mean or average horsepower** of the cars among the population of cars, we will check and calculate the average of all values. In this case, we’ll take the sum of the Horse power of each car, divided by the total number of cars:

**Mean = (110+110+93+96+90+110+110+110)/8 = 103.625**

If we want to find out the **center value of mpg** among the population of cars, we will arrange the mpg values in ascending or descending order and choose the middle value. In this case, we have 8 values which is an even entry. Hence we must take the average of the two middle values.

The **mpg for 8 cars**: 21,21,21.3,22.8,23,23,23,23

**Median** **= (22.8+23) /2 = 22.9**

If we want to find out the most common type of cylinder among the population of cars, we will check the value which is repeated the most number of times. Here we can see that the cylinders come in two values, 4 and 6. Take a look at the data set; you can see that the most recurring value is 6. Hence **6 is our Mode.**

* 1. **Discrete Observations:**

**Ex1:** Monthly sales (in 2021 *₹) of 10 small shops are given below.*

*100,190,210,160,150,160,190,200,170,152 . Calculate mean, mode and median*

**Solution: R- Code with output**

|  |
| --- |
| > x=c(100,190,210,160,150,160,190,200,170,152)  > n=length(x)  > n  [1] 10  > tx=table(x)  > tx  x  100 150 152 160 170 190 200 210  1 1 1 2 1 2 1 1  > m=which(tx==max(tx)) #which()returns the position of the specified values in the logical vector.  > m  160 190  4 6  > stx=sort(unique(x)) #unique() eliminates or delete the duplicate values or the rows present  > stx  [1] 100 150 152 160 170 190 200 210  mo=stx[m]  > cat("Mode=",mo,"\n");  Mode= 160 190  > cat("Median=",me,"\n");  Median= 165  > am=mean(x)  > cat("mean=",am,"\n");  mean= 168.2 |

Each of above cat command displays corresponding value

* 1. **Ungrouped Frequency Distribution**

Let (xi,fi),i=1,2,….,k be a given ungrouped frequency distribution, then we use the following formulae to calculate different measures.

**Mean**=

**Mode** = The observation with highest frequency in the data

Median = Value of the middlemost observation, when the observation arranged in the order.

**Ex2:** For the following frequency distribution

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **x** | **1** | **2** | **3** | **4** | **5** |
| **f** | **7** | **11** | **9** | **8** | **3** |

Calculate Mean, Median and Mode

**Solution: R- Code with output**

|  |
| --- |
| > x=1:5;f=c(7,11,9,8,3);n=sum(f);  > y=rep(x,f)#rep allows you to repeat a scalar (or vector) a specified number of times, or to a desired length.  > y  [1] 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 5 5 5  > am=mean(y)  > m=which(f==max(f));  > mo=x[m];  > me=median(y);  > cat("Mean=",am,"\n");  Mean= 2.710526  > cat("Mode=",mo,"\n");  Mode= 2  > cat("Median=",me,"\n");  Median= 3 |

* 1. **Grouped Frequency Distribution**

For a grouped frequency distribution, let Xi denote the class mark of the ith class and let fi be the frequency of the ith class i=1,2,…k then we use the following formulae to calculate different measures.

**Mean**=

**Median**=L+(N/2-CF)\*(h/f)

Where,

* L = lower limit of median class
* N = total number of observations
* CF = cumulative frequency of the preceding class
* f = frequency of median class
* h = class size (upper limit - lower limit)

**Mode**=L+((fm-f1)/(2fm-f1-f2))\*h

where,

* L is the lower limit of the modal class
* h is the size of the class interval
* fm is the frequency of the modal class
* f1 is the frequency of the class preceding the modal class
* f2 is the frequency of the class succeeding the modal class

***Note:*** *The observation with the highest frequency is the modal value for the given data is here referred to as the* ***modal value****.*

**Ex3:** The frequency distribution of weight (in grams) of mangoes of a certain variety is given below.

|  |  |
| --- | --- |
| **Weight** | **No. of Mangoes** |
| 410-420 | 14 |
| 420-430 | 20 |
| 430-440 | 42 |
| 440-450 | 54 |
| 450-460 | 45 |
| 460-470 | 18 |
| 470-480 | 7 |

Calculate mean, mode and median.

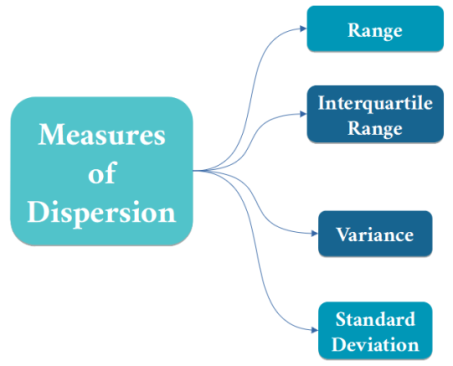
**Solution: R- Code with output**

|  |
| --- |
| > lb=seq(410,470,10);  > ub=seq(420,480,10); h=10;  > f=c(14,20,42,54,45,18,7);  > x=(lb+ub)/2; n=sum(f);  > lb  [1] 410 420 430 440 450 460 470  > ub  [1] 420 430 440 450 460 470 480  > x  [1] 415 425 435 445 455 465 475  > n  [1] 200  > lcf=cumsum(f); #cumsum()calculates the cumulative sum of the vector passed as argument.  > lcf  [1] 14 34 76 130 175 193 200  > mc=min(which(lcf>=n/2));  > mc  [1] 4  > med=lb[mc]+(n/2-lcf[mc-1])\*(h/f[mc]);  > moc=which(f==max(f));  > mo=lb[moc]+((f[moc]-f[moc-1])/(2\*f[moc]-f[moc-1]-f[moc+1]))\*h;  > mean=sum(f\*x)/n;  > cat("Mean=",mean,"\n");  Mean= 443.9  > cat("Mode =",mo,"\n");  Mode = 445.7143  > cat("Median =",med,"\n");  Median = 444.4444 |

***Note:*** *Cumulative sums, or running totals, are used to display the total sum of data as it grows with time (or any other series or progression).*

1. **Measures of Spread (Measure of Dispersion)**

Measures of spread (sometimes also called a measure of dispersion) describe how similar or varied the set of observed values are for a particular variable (data item). Measures of spread include the range, quartiles and the interquartile range, variance and standard deviation.



The spread of the values can be measured for quantitative data, as the variables are numeric and can be arranged into a logical order with a low end value and a high end value.

## Reasons to measure spread

Summarizing the dataset can help us understand the data, especially when the dataset is large. As discussed in the Measures of central tendency page, the mode, median, and mean summarize the data into a single value that is typical or representative of all the values in the dataset, but this is only part of the 'picture' that summarizes a dataset.   Measures of spread summarize the data in a way that shows how scattered the values are and how much they differ from the mean value.

| **Dataset examples** | |
| --- | --- |
| **Dataset A** | **Dataset B** |
| 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8 | 1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11 |

The mode (most frequent value), median (middle value\*) and mean (arithmetic average) of both datasets is 6.

(\*note, the median of an even numbered data set is calculated by taking the mean of the middle two observations).

If we just looked at the measures of central tendency, we may assume that the datasets are the same.

However, if we look at the spread of the values in the following graph, we can see that Dataset B is more dispersed than Dataset A. Used together, the measures of central tendency and measures of spread help us to better understand the data.

|  |  |
| --- | --- |
| **Spread of values in Dataset A and Dataset B** | |
| Column graph comparing the spread of values in dataset A and dataset B | * 0 - Dataset A = 0, Dataset B = 0 * 1 - Dataset A = 0, Dataset B = 1 * 2 - Dataset A = 0, Dataset B = 1 * 3 - Dataset A = 0, Dataset B = 1 * 4 - Dataset A = 1, Dataset B = 1 * 5 - Dataset A = 3, Dataset B = 1 * 6 - Dataset A = 4, Dataset B = 2 * 7 - Dataset A = 3, Dataset B = 1 * 8 - Dataset A = 1, Dataset B = 1 * 9 - Dataset A = 0, Dataset B = 1 * 10 - Dataset A = 0, Dataset B = 1 * 11 - Dataset A = 0, Dataset B = 1 * 12 - Dataset A = 0, Dataset B = 0 |

Just like the measure of center, we also have measures of the spread, which comprises of the following measures:

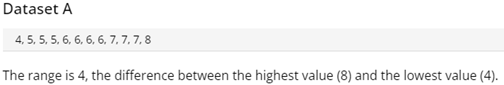
### Range

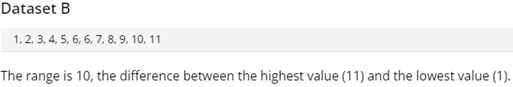
The range is the difference between the smallest value and the largest value in a dataset.

**Range = Max(x\_i) – Min(x\_i)**Here,

Max(x\_i): Maximum value of x

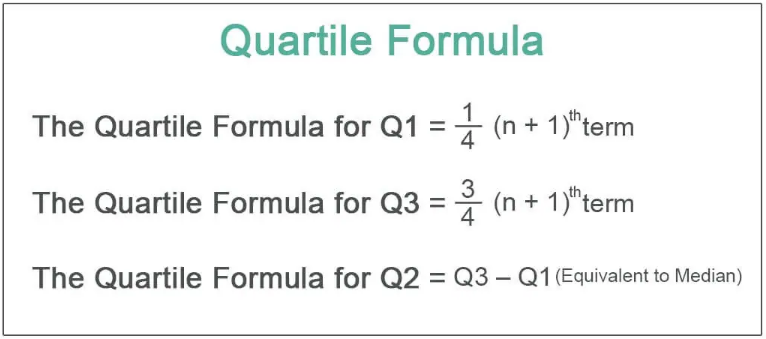
Min(x\_i): Minimum value of x





### Quartiles

Quartiles divide an ordered dataset into four equal parts, and refer to the values of the point between the quarters. A dataset may also be divided into quintiles (five equal parts) or deciles (ten equal parts).



To better understand how quartile is calculated, let’s look at an example.

To illustrate the steps for how to find quartiles, consider the scores on the previously mentioned test.

**Step 1**

List the data points in numerical order from lowest to highest, if they are not already in order.

|  |
| --- |
|  |
| ***The first step to finding the Interquartile range is to list a set of data in numerical order.*** |

**Step 2**

Identify the median of the data set by finding the data point in the exact middle of the set. For an odd number of data points this simply entails identifying the datum that has an equal number of points to its left and right.

|  |
| --- |
|  |
| ***Finding the median is imperative before either the first quartile or third quartile can be found.*** |

**Step 3**

Next, find the middle data point within the left half (bottom) of the data and the middle data point within the right half (top) of the data. Note that the median is not included in either the bottom or the top portions of the data set.

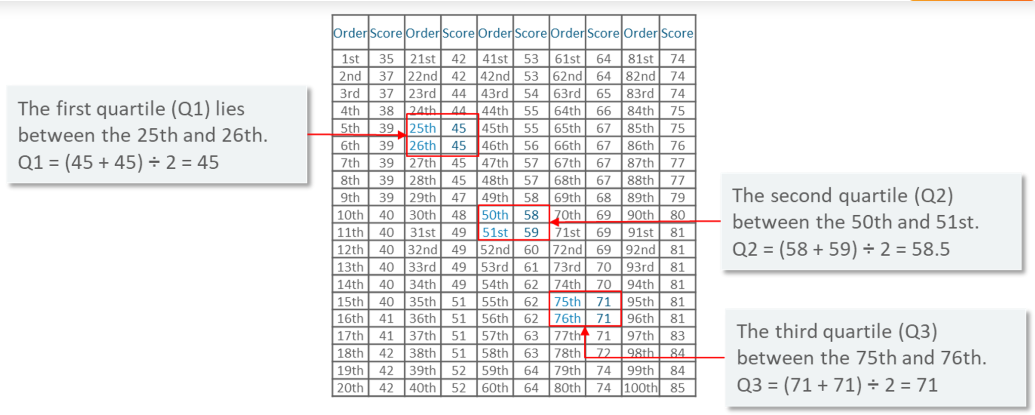
|  |
| --- |
|  |
| ***The first quartile is halfway between the lowest score and the median, while the third quartile is located halfway from the median to the top of the data set.*** |

**Step 4**

Label the first quartile, **Q1**, and the third quartile, **Q3**.



|  |
| --- |
| ***Q1 and Q3 are used to label the first and third, respectively, quartiles.***  ***Example:*** |



The above image shows marks of 100 students ordered from lowest to highest scores. The quartiles lie in the following ranges:

1. The first quartile (Q1) lies between the 25th and 26th observation.
2. The second quartile (Q2) lies between the 50th and 51st observation.
3. The third quartile (Q3) lies between the 75th and 76th observation.
4. **Inter Quartile Range (IQR)**

It is the measure of variability, based on dividing a data set into quartiles. The interquartile range is equal to Q3 minus Q1,

i.e. **IQR = Q3 – Q1**Quartile Deviation =**( Q3 – Q1)/2**

1. **Variance**

The term variance refers to a statistical measurement of the spread between numbers in a data set. More specifically, variance measures how far each number in the set is from the [mean](https://www.investopedia.com/terms/m/mean.asp) (average), and thus from every other number in the set.

*Variance* is average squared distance from mean (squared because certain data points will be less than mean and will negate the values when added).Variance can be calculated by using the below formula:

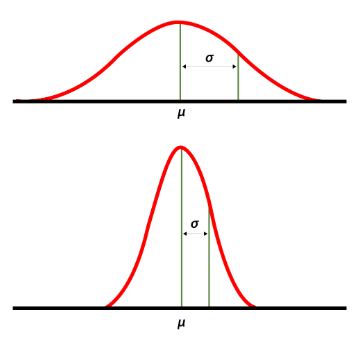
σ**2**=

Here,

x: Individual data points  
n: Total number of data points  
x̅: Mean of data points

1. **Standard Deviation**

A **standard deviation** (or σ) is a measure of how dispersed the data is in relation to the mean. Low, or small, standard deviation indicates data are clustered tightly around the mean, and high, or large, standard deviation indicates data are more spread out. A standard deviation close to zero indicates that data points are very close to the mean, whereas a larger standard deviation indicates data points are spread further away from the mean.



In the image, the curve on top is more spread out and therefore has a higher standard deviation, while the curve below is more clustered around the mean and therefore has a lower standard deviation.

To calculate the standard deviation, use the following formula:

σ=

Where ,

σ is the standard deviation,

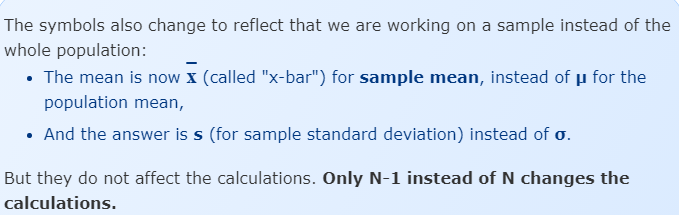
xi is each individual data point in the set,

µ is the mean,

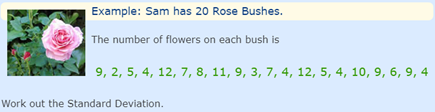
N is the total number of data points.

The Population Standard Deviation: σ =

The Sample Standard Deviation: s =



**Example**



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rose bushes** | **No. of Flowers**  **(X)** | **µ (Mean)** | **Absolute deviation from Mean**  (xi -µ)2 |  | **Variance** |
| 1 | 9 | 140/20 **=7** | (9-7)2=4 | **178** | (1/20)\*178  **=8.9** |
| 2 | 2 | 25 |
| 3 | 5 | 4 |
| 4 | 4 | 9 |
| 5 | 12 | 25 |
| 6 | 7 | 0 |
| 7 | 8 | 1 |
| 8 | 11 | 16 |
| 9 | 9 | 4 |
| 10 | 3 | 16 |
| 11 | 7 | 0 |
| 12 | 4 | 9 |
| 13 | 12 | 25 |
| 14 | 5 | 4 |
| 15 | 4 | 9 |
| 16 | 10 | 9 |
| 17 | 9 | 4 |
| 18 | 6 | 1 |
| 19 | 9 | 4 |
| 20 | 4 | 9 |

**Standard Deviation:**

σ =

σ = =2.983

**Measure of Dispersion Formulas**

|  |  |  |
| --- | --- | --- |
| **Sl.No.** | **Measure** | **Formula** |
| 1. | Range | Range=Max – Min |
| 2. | Coefficient of Range | Cr=(Max-Min)/(Max+Min) |
| 3. | Variance | σ**2**= |
| 4. | Standard Deviation | σ = |
| 5. | Coefficient of Variance | σ/|Mean| |
| 5. | Mean Deviation from Mean |  |
| 6. | Mean Deviation about median |  |
| 6. | Quartile Deviation | (Q3-Q1)/2 |
| 7. | Coefficient of quartile deviation | (Q3-Q1)/(Q3+Q1) |

## ****Ex1:** The number of mistakes in a page recorded for 20 pages are as follows**

## **2,5,9,7,11,6,5,2,7,9,3,2,8,12,14,6,3,9,8,7.**

## **Find *(i) Range and Coefficient of Range (ii) Quartile Deviation and Coefficient of quartile deviation (iii) Mean deviation from mean (iv) Coefficient of variance***

## ****Solution: R-Code with output****

|  |
| --- |
| > x=c(2,5,9,7,11,6,5,2,7,9,3,2,8,12,14,6,3,9,8,7);  > n=length(x);  > mx=mean(x);q1=quantile(x,0.25);q3=quantile(x,0.75);  > mi=min(x);ma=max(x);r=ma-mi;cr=r/(ma+mi);qd=(q3-q1)/2;  > vl=var(x); #variance function uses denominator(n-1)  > v=((n-1)/n)\*vl;  > sd=v^0.5;  > cv=sd\*100/abs(mx) #coefficient of variance in percentage  > md=sum(abs(x-mx))/n  > cat("Mean =",mx,"\n");  Mean = 6.75  > cat("range=",r,"\n");  range= 12  > cat("coefficient of range=",cr,"\n");  coefficient of range= 0.75  > cat("quartile deviation=",qd,"\n");  quartile deviation= 2.25  > cat("coefficient of variation=",cv,"\n");  coefficient of variation= 49.10726  > cat("Mean deviation from mean=",md,"\n");  Mean deviation from mean= 2.675 |

## ****Ex2:** Find (i) quartile deviation and coefficient of quartile deviation (ii) Mean deviation about median (iii) coefficient of variation for the following data**

|  |  |
| --- | --- |
| ****No. of goals scored in a match**** | ****No. of matches**** |
| **0** | **25** |
| **1** | **9** |
| **2** | **8** |
| **3** | **5** |
| **4** | **4** |

## ****Solution: R-Code with output****

|  |
| --- |
| > x=0:4;f=c(25,9,8,5,4);n=sum(f);y=rep(x,f);*#rep repeats a vector x*  > mx=sum(f\*x)/n;q1=quantile(y,0.25);q3=quantile(y,0.75);  > me=quantile(y,0.5);  > qd=(q3-q1)/2;  > cqd=(q3-q1)/(q3+q1);  > v=sum(f\*(x-mx)^2)/n;  > sd=v^0.5;  > cv=sd\*100/abs(mx);  > md=sum(f\*abs(x-me))/n;  > cat("Mean=",mx,"\n");  **Mean= 1.098039**  > cat("Median=",me,"\n");  **Median= 1**  > cat("Quartile deviation=",qd,"\n");  **Quartile deviation= 1**  > cat("coefficient of Quartile deviation=",cqd,"\n");  **coefficient of Quartile deviation= 1**  > cat("coefficient of variation=",cv,"\n");  **coefficient of variation= 119.9756**  > cat("Mean about median=",md,"\n");  **Mean about median= 1.078431** |

## ****Quartile formula for grouped data****

Qi = L + (i\*N/4 – M)\* (C/F)

*Where,*

* *L: The lower bound of the interval that contains the ith quartile*
* *C: The class width*
* *F: The frequency of the interval that contains the ithquartile*
* *N: The total frequency*
* *M: The* cumulative frequency leading up to the interval that contains the ith quartile

## ****Ex3.** For the following frequency distribution of size of holdings, calculate *(i) quartile deviation (ii) standard deviation (iii) coefficient of variation***

|  |  |
| --- | --- |
| ****Size of holdings (hectares)**** | ****No. of farms**** |
| **2.5-3.5** | **1000** |
| **3.5-4.5** | **2300** |
| **4.5-5.5** | **3600** |
| **5.5-6.5** | **2400** |
| **6.5-7.5** | **1700** |
| **7.5-8.5** | **1200** |
| **8.5-9.5** | **520** |

## ****Solution: R-Code with output****

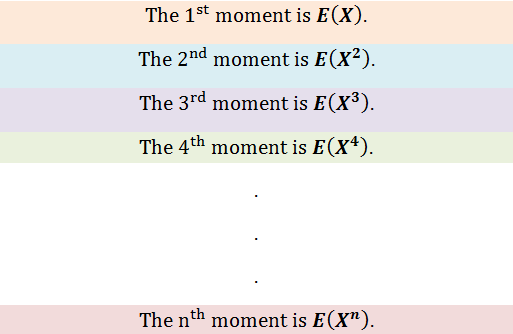
|  |
| --- |
| > lb=2.5:8.5;ub=3.5:9.5;h=1; #h is class width  > f=c(1000,2300,3600,2400,1700,1200,520);  > x=(lb+ub)/2;  > n=sum(f);  > am=sum(f\*x)/n;  > v=sum(f\*(x-am)^2)/n;  > sd=v^0.5;  > cv=sd\*100/abs(am);  > lcf=cumsum(f);  > lcf;  **[1] 1000 3300 6900 9300 11000 12200 12720**  > n;  **[1] 12720**  > n/4;  **[1] 3180**  > 3\*n/4;  **[1] 9540**  > q1c=min(which(lcf>=n/4));  > q3c=min(which(lcf>=3\*n/4));  > q3c  **[1] 5**  > lb[q3c];  **[1] 6.5**  > lcf[q3c-1];  **[1] 9300**  > f[q3c];  **[1] 1700**  > q3=lb[q3c]+(3\*n/4-lcf[q3c-1])\*(h/f[q3c]);  > q1=lb[q1c]+(1\*n/4-lcf[q1c-1])\*(h/f[q1c]);  > qd=(q3-q1)/2;  > cat("quartile deviation=",qd,"\n");  **quartile deviation= 1.096675**  > cat("standard Deviation=",sd,"\n");  **standard Deviation= 1.550186**  > cat("coefficient of variation=",cv,"\n");  **coefficient of variation= 27.85867** |

## ****Moments, Skewness and Kurtosis****

**What is a Moment in Statistics?**

We generally use moments in statistics, machine learning, mathematics, and other fields to describe the characteristics of a distribution.

Let’s say the variable of our interest is X then, moments are X’s expected values. For example, E(X), E(X²), E(X³), E(X⁴),…, etc.



*Fig: Moments in statistics*

**Moments in statistics:**

1) **First Moment:** Measure of the central location. **(MEAN)**

2) **Second Moment:** Measure of dispersion/spread.**(VARIANCE)**

3) **Third Moment:** Measure of asymmetry.

4) **Fourth Moment:** Measure of outliers/tailedness.

Now we are very familiar with the first moment (mean) and the second moment(variance).

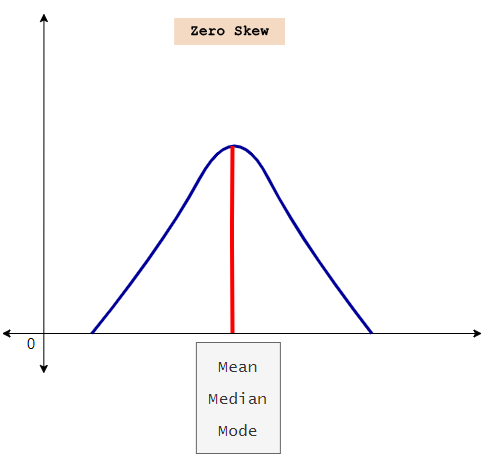
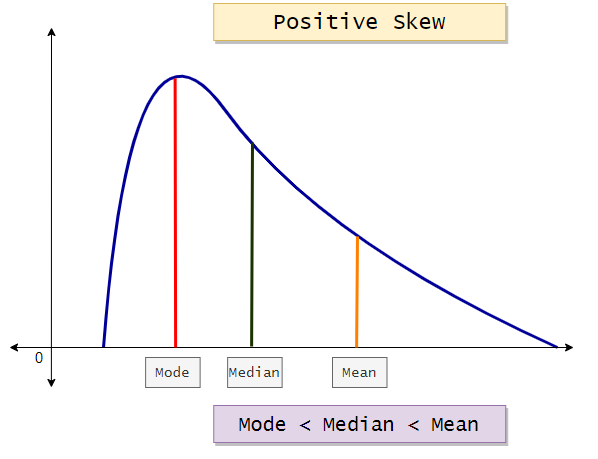
The third moment is called **skewness**, and the fourth moment is known as **kurtosis.**

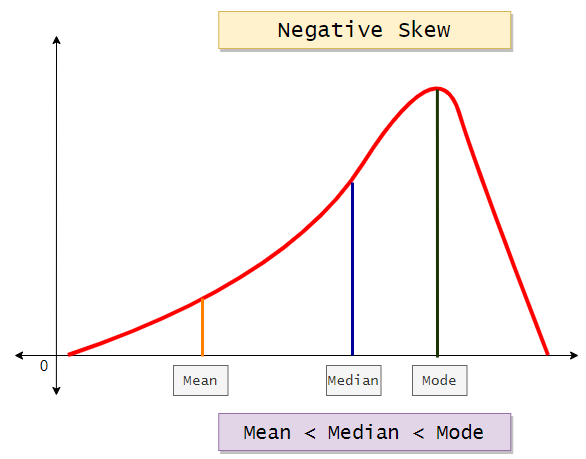
The third moment measures the asymmetry of distribution while the fourth moment measures how heavy the tail values are. Physicists generally use the higher-order moments in applications of physics.

**Skewness**

Skewness is a measure of the asymmetry of a distribution. A distribution is asymmetrical when its left and right side are not mirror images.

A distribution can have *right (or positive), left (or negative), or zero skewness*. A right-skewed distribution is longer on the right side of its peak, and a left-skewed distribution is longer on the left side of its peak:

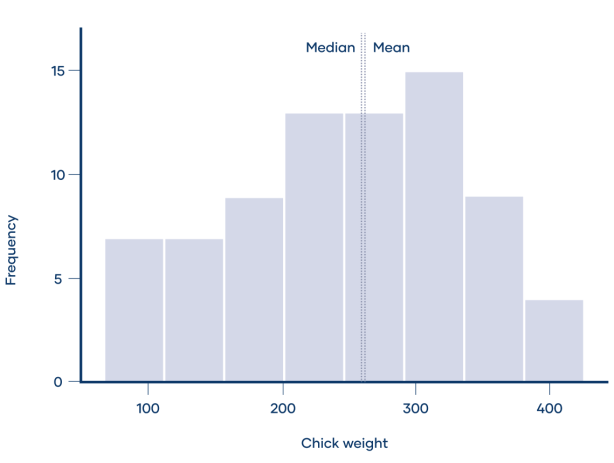


**What is zero skew?**

When a distribution has zero skew, it is symmetrical. Its left and right sides are mirror images.

Normal distributions have zero skew, but they’re not the only distributions with zero skew. Any symmetrical distribution, such as a uniform distribution or some bimodal (two-peak) distributions, will also have zero skew. The easiest way to check if a variable has a skewed distribution is to plot it in a histogram. For example, the weights of six-week-old chicks are shown in the histogram below.

The distribution is approximately symmetrical, with the observations distributed similarly on the left and right sides of its peak. Therefore, the distribution has approximately zero skew.



In a distribution with zero skew, the mean and median are equal.

Zero skew: mean = median

For example, the mean chick weight is 261.3 g, and the median is 258 g. The mean and median are almost equal. They aren’t perfectly equal because the sample distribution has a very small skew.

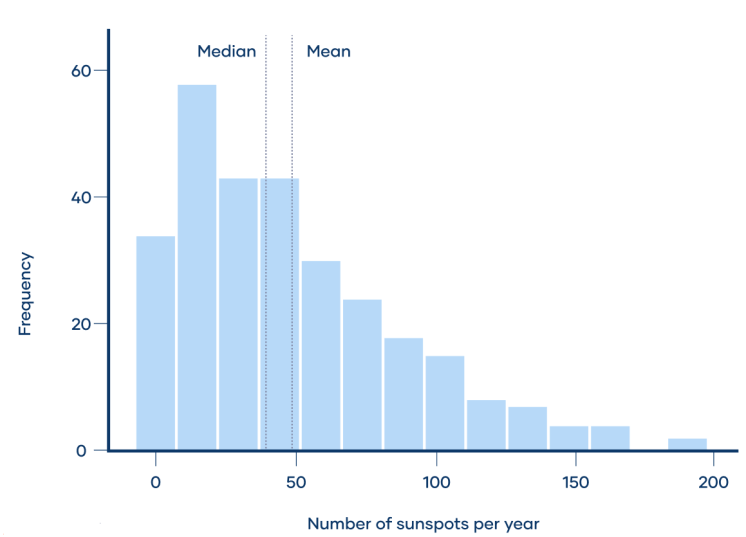
**What is right skew (positive skew)?**

A right-skewed distribution is longer on the right side of its peak than on its left. Right skew is also referred to as positive skew.

Skewness is determined in terms of tails. A tail is a long, tapering end of a distribution. It indicates that there are observations at one of the extreme ends of the distribution, but that they’re relatively infrequent. A right-skewed distribution has a long tail on its right side.

The number of sunspots observed per year, shown in the histogram below, is an example of a right-skewed distribution. The sunspots, which are dark, cooler areas on the surface of the sun, were observed by astronomers between 1749 and 1983.

The distribution is right-skewed because it’s longer on the right side of its peak. There is a long tail on the right, meaning that every few decades there is a year when the number of sunspots observed is a lot higher than average.



The mean of a right-skewed distribution is almost always greater than its median. That’s because extreme values (the values in the tail) affect the mean more than the median.

**Right skew: mean > median**

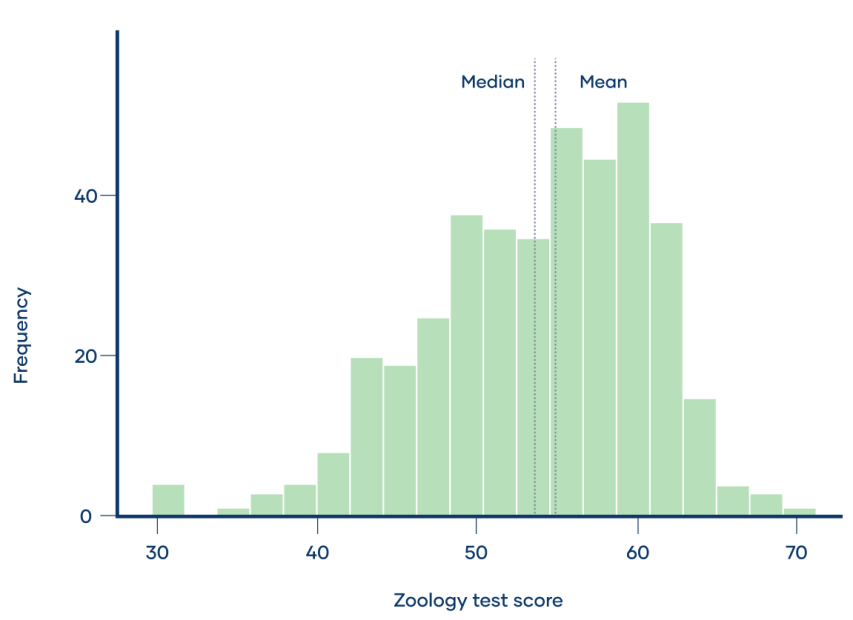
For example, the mean number of sunspots observed per year was 48.6, which is greater than the median of 39.

**What is left skew (negative skew)?**

A left-skewed distribution is longer on the left side of its peak than on its right. In other words, a left-skewed distribution has a long tail on its left side. Left skew is also referred to as negative skew.

Test scores often follow a left-skewed distribution, with most students performing relatively well and a few students performing far below average. The histogram below shows scores for the zoology portion of a standardized test taken by Indian students at the end of high school.

The distribution is left-skewed because it’s longer on the left side of its peak. The long tail on its left represents the small proportion of students who received very low scores.



The mean of a left-skewed distribution is almost always less than its median.

**Left skew: mean < median**

For example, the mean zoology test score was 53.7, which is less than the median of 55.

**Formula to calculate skewness**

**Bowley coefficient of skewness**

**Example:** Find the Bowley skewness for the following set of data

|  |  |  |
| --- | --- | --- |
| **# of pets** | **# of families** | **Cumulative freq** |
| 0 | 60 | 60 |
| 1 | 60 | 120 |
| 2 | 5 | 170 |
| 3 | 20 | 190 |
| 4 | 25 | 215 |
| 5 | 10 | 225 |
| 6 or more | 5 | 230 |

**Step1:** Find the quartile for the data set.

Q1 = (total cum freq+1/4)th observation= (230+1/4) =57.75

Q2 = (total cum freq+1/2)th observation= (230+1/2) =115.5

Q3 = 3(total cum freq+1/4)th observation= 3(230+1/4) =173.25

**Step 2:** Look into the table to find the nth observations you calculated in step1.

Q1 = 57.7th observation=0

Q2 = 115.5th observation=1

Q3=173.25th observation=3

**Step 3:** Plug the above values into the formula

Skq = (Q3+Q1-2Q2)/(Q3-Q1)

Skq = 3+0-2/3-0=1/3

Skq = +1/3, so the distribution is positively skewed.

|  |  |  |
| --- | --- | --- |
| **Karl Pearson’s coefficient of skewness** | **=** | It is used when mean, median, mode and standard deviation is known |
| **Bowley coefficient of skewness** | (Q3+Q1-2Q2)/(Q3-Q1) | It is used if we have only quartile information |

**Ex1:** For the following observations.

10,12,14,16,12,8,23,10,12,20,30,25,27,12

Find *(i) Karl pearson’s coefficient of skewness (ii) Bowley’s coefficient of skewness*

**Solution: R code with output**

|  |
| --- |
| > x=c(10,12,14,16,12,8,23,10,12,20,30,25,27,12);  > n=length(x);  > tx=table(x);  > stx=sort(unique(x));  > mo=stx[m];  > mx=mean(x);  > v=((n-1)/n)\*var(x);  > sd=v^0.5  > skp=(mx-mo)/sd;  > q1=quantile(x,0.25);  > q2=quantile(x,0.5);  > q3=quantile(x,0.75);  > skb=(q3+q1-2\*q2)/(q3-q1);  > cat("Mean =",mx,"\n");  **Mean = 16.5**  > cat("Mode =",mo,"\n");  **Mode = 12**  > cat("Standard Deviation =",sd,"\n");  **Standard Deviation = 6.884247**  > cat("Karl Pearson's coefficient=",skp,"\n");  **Karl Pearson's coefficient= 0.6536663**  > cat("Bowley's coefficient=",skb,"\n");  **Bowley's coefficient= 0.804878** |

**Ex2.** Frequency distribution of number of absent days in a week of college students is given below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| No. of absent days | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| No. of students | 3 | 9 | 17 | 23 | 10 | 7 | 1 |

*Calculate (i) Karl Pearson’s coefficient of skewness (ii) Bowley’s coefficient of skewness*

**Solution: R code with output**

|  |
| --- |
| > x=0:6;f=c(3,9,17,23,10,7,1); n=sum(f);  > m=which(f==max(f));  > mo=x[m];  > y=rep(x,f);  > mx=mean(y);  > v=((n-1)/n)\*var(y);  > sd=v^0.5;  > skp=(mx-mo)/sd;  > q1=quantile(y,0.25);  > q2=quantile(y,0.5);  > q3=quantile(y,0.75);  > skb=(q3+q1-2\*q2)/(q3-q1);  > cat("Mean=",mx,"\n");  **Mean= 2.757143**  > cat("Mode=",mo,"\n");  **Mode= 3**  > cat("Standard Deviation=",sd,"\n");  **Standard Deviation= 1.32488**  > cat("Karl Pearson's coefficient of skewness=",skp,"\n");  **Karl Pearson's coefficient of skewness= -0.1833051**  > cat("Bowley's coefficient of skewness=",skb,"\n");  **Bowley's coefficient of skewness= -0.1428571** |

**Ex.3: Frequency distribution of number of persons according to their monthly expenditure on transport is given below.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Monthly Expenditure** | **300-350** | **350-400** | **400-450** | **450-500** | **500-550** | **550-600** |
| **No. of persons** | **12** | **23** | **38** | **15** | **10** | **2** |

***Calculate (i) Karl Pearson’s coefficient of skewness (ii) Bowley’s coefficient of skewness***

**Solution: R code with output**

|  |
| --- |
| **> lb=seq(300,550,50);ub=seq(350,600,50);**  **> f=c(12,23,38,15,10,2);x=(lb+ub)/2;n=sum(f);**  **> mx=sum(f\*x)/n;v=sum(f\*(x-mx)^2)/n;sd=v^0.5;h=50;**  **> moc=which(f==max(f));**  **> mo=lb[moc]+((f[moc]-f[moc-1])/(2\*f[moc]-f[moc-1]-f[moc+1]))\*h;**  **> skp=(mx-mo)/sd;**  **> lcf=cumsum(f);**  **> q1c=min(which(lcf>=n/4));**  **> q2c=min(which(lcf>=n/2));**  **> q3c=min(which(lcf>=3\*n/4));**  **> q3=lb[q3c]+(3\*n/4-lcf[q3c-1])\*(h/f[q3c]);**  **> q2=lb[q2c]+(n/2-lcf[q2c-1])\*(h/f[q2c]);**  **> q1=lb[q1c]+(n/4-lcf[q1c-1])\*(h/f[q1c]);**  **> skb=(q3+q1-2\*q2)/(q3-q1);**  **> cat("Mean=",mx,"\n");**  **Mean= 422**  **> cat("Mode=",mo,"\n");**  **Mode= 419.7368**  **> cat("Standard Deviation=",sd,"\n");**  **Standard Deviation= 59.92495**  **> cat("Karl Pearson's coefficient of skewness=",skp,"\n");**  **Karl Pearson's coefficient of skewness= 0.03776654**  **> cat("Bowley's coefficient of skewness=",skb,"\n");**  **Bowley's coefficient of skewness= -0.05798229** |

**Finding skewness using library functions**

**Install the package *moments* and load the *moments* package in R**

**Solution: R code with output**

|  |
| --- |
| **> data=c(10,12,14,16,12,8,23,10,12,20,30,25,27,12);**  **> skewness(data);**  **[1] 0.6580996**  **> hist(data, col='steelblue')** |

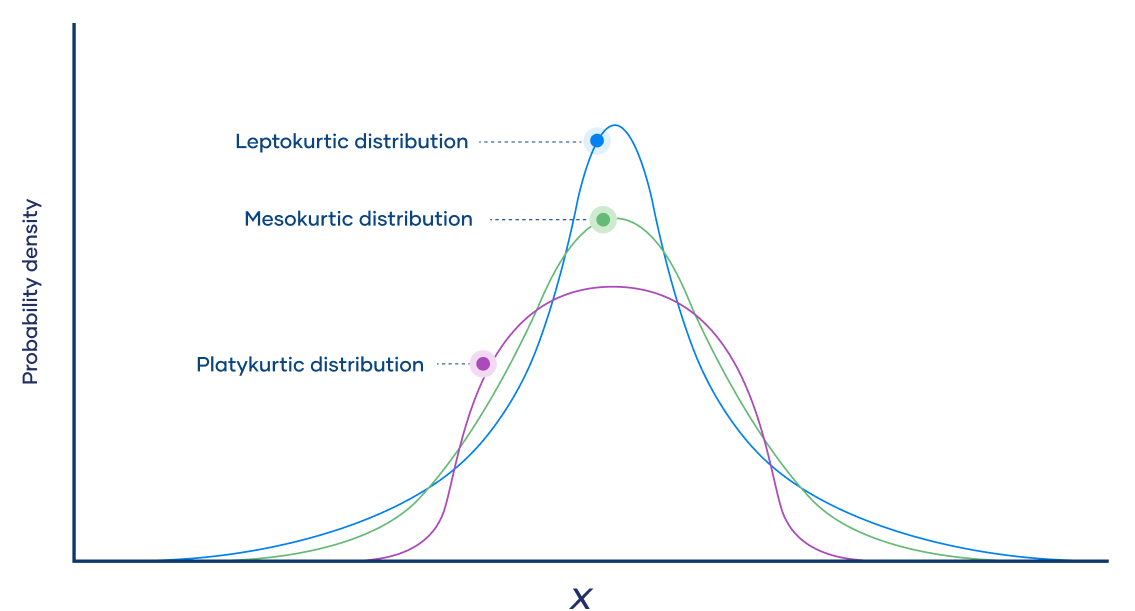
**Kurtosis**

Kurtosis is a measure used to describe the distribution of observed data around the mean. Excess kurtosis is the tailedness of a distribution relative to a normal distribution.

* Distributions with medium kurtosis (medium tails) are [**mesokurtic**](https://www.scribbr.com/statistics/kurtosis/#mesokurtic)**.**
* Distributions with low kurtosis (thin tails) are [**platykurtic**](https://www.scribbr.com/statistics/kurtosis/#platykurtic)**.**
* Distributions with high kurtosis (fat tails) are [**leptokurtic**](https://www.scribbr.com/statistics/kurtosis/#leptokurtic).

*Tails*are the tapering ends on either side of a distribution. They represent the probability or frequency of values that are extremely high or low compared to the mean. In other words, tails represent how often outliers occur.

**Example: Types of kurtosis**



A **mesokurtic** distribution is medium-tailed, so outliers are neither highly frequent, nor highly infrequent.

A **platykurtic** distribution is thin-tailed, meaning that outliers are infrequent.

A **leptokurtic** distribution is fat-tailed, meaning that there are a lot of outliers.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Category** | | |
| **Mesokurtic** | **Platykurtic** | **Lepokurtic** |
| **Taildness** | Medium-tailed | Thin-tailed | Fat-tailed |
| **Outlier frequency** | Medium | Low | High |
| **Kurtosis** | Moderate(3) | Low(<3) | High(>3) |
| **Excess kurtosis** | 0 | Negative | Positive |
| **Example** | Normal | Uniform | Laplace |

**Kurtosis of a population**

The following formula describes the kurtosis of a population:

**Kurtosis = µ4/σ2**

Where:

* µ4 is the unstandardized central fourth moment
* σ2 is the standard deviation

**Excess Kurtosis = Kurtosis - 3**

**Ex1.** For the following observations

3,6,3,8,2,9,4,2,10,11,2,4

*Compute (i) First four central moments (ii) Measure of kurtosis based on moments*

**Solution: R code with output**

|  |
| --- |
| > x=c(3,6,3,8,2,9,4,2,10,11,2,4);n=length(x);  > mx=mean(x);  > cm2=sum((x-mx)^2)/n;  > cm3=sum((x-mx)^3)/n;  > cm4=sum((x-mx)^4)/n;  > kurtosis=cm4/(cm2^2);  > EK=kurtosis-3;  > cat("second central moment=",cm2,"\n");  **second central moment= 10.22222**  > cat("Third central moment=",cm3,"\n");  **Third central moment= 17.57407**  > cat("Fourth central moment=",cm4,"\n");  **Fourth central moment= 181.0741**  > cat("Measure of kurtosis based on moments=",kurtosis,"\n");  **Measure of kurtosis based on moments= 1.732869**  > cat("Measure of Excess kurtosis based on moments=",EK,"\n");  **Measure of Excess kurtosis based on moments= -1.267131** |

**Finding kurtosis using library functions**

**Install the package *moments* and load the *moments* package in R**

**Solution: R code with output**

|  |
| --- |
| > data=c(3,6,3,8,2,9,4,2,10,11,2,4);  > kurtosis(data);  **[1] 1.732869** |

**Ex.2** Frequency distribution of number of attempts required to pass the final examination is given below

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **No. of Attempts** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| **No. of Students** | 12 | 23 | 45 | 32 | 12 | 5 | 2 |

Compute *(i) First four central moments (ii) Measure of kurtosis based on moments*

**Solution: R code with output**

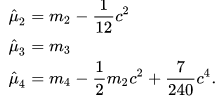
|  |
| --- |
| > x=1:7; f=c(12,23,45,32,13,5,2);  > n=sum(f);  > mx=sum(f\*x)/n;  > cm2=sum((x-mx)^2)/n;  > cm3=sum((x-mx)^3)/n;  > cm4=sum((x-mx)^4)/n;  > kurtosis=cm4/cm2^2;  > EK=kurtosis-3;  > cat("second central moment=",cm2,"\n");  **second central moment= 0.2413512**  > cat("Third central moment=",cm3,"\n");  **Third central moment= 0.4941528**  > cat("Fourth central moment=",cm4,"\n");  **Fourth central moment= 2.202479**  > cat("Measure of kurtosis based on moments=",kurtosis,"\n");  **Measure of kurtosis based on moments= 37.81055**  > cat("Measure of Excess kurtosis based on moments=",EK,"\n");  **Measure of Excess kurtosis based on moments= 34.81055** |

**Finding kurtosis using library functions**

|  |
| --- |
| > # Defining data vector  > x <-c(rep(1,each=12),rep(2,each =23),rep(3,each =45),rep(4,each =32),rep(5,each =12),rep(6,each =5),rep(7,each =2));  > kurtosis(x);  **[1] 3.162707** |

**Sheppard's correction**

Sheppard's corrections are approximate corrections to estimates of moments computed from binned data (deleted data).

****

**Ex3.** Frequency distribution of duration of advertisements on television is as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Duration(in seconds)** | 30-35 | 35-40 | 40-45 | 45-50 | 50-55 | 55-60 |
| **No. of Advertisements** | 6 | 18 | 25 | 12 | 5 | 3 |

*Compute (i) first four central moments (ii) Measure of kurtosis based on moments*

**Solution: R code with output**

|  |
| --- |
| > lb=seq(30,55,5);ub=seq(35,60,5);  > f=c(6,18,25,12,5,3);  > x=(lb+ub)/2;  > n=sum(f);h=5;  > mx=sum(f\*x)/n;  > cm2=sum((x-mx)^2)/n;  > cm3=sum((x-mx)^3)/n;  > cm4=sum((x-mx)^4)/n;  > ccm2=cm2-h^2/12; #corrected second central moment  > ccm4=cm4-(h^2/2)\*cm2+(7/240)\*h^4; #corrected second central moment  > kurtosis=ccm4/ccm2^2;  > EK=kurtosis-3;  > cat("second central moment=",cm2,"\n");  **second central moment= 6.853009**  > cat("Corrected second central moment=",ccm2,"\n");  **Corrected second central moment= 4.769675**  > cat("Third central moment=",cm3,"\n");  **Third central moment= 47.4199**  > cat("Fourth central moment=",cm4,"\n");  **Fourth central moment= 1027.706**  > cat("Corrected Fourth central moment=",ccm4,"\n");  **Corrected Fourth central moment= 960.2721**  > cat("Measure of kurtosis based on moments=",kurtosis,"\n");  **Measure of kurtosis based on moments= 42.21013**  > cat("Measure of Excess kurtosis based on moments=",EK,"\n");  **Measure of Excess kurtosis based on moments= 39.21013** |

## ****Probability****

Probability is an intuitive concept. We use it on a daily basis without necessarily realizing that we are speaking and applying probability to work.

Life is full of uncertainties. We don’t know the outcomes of a particular situation until it happens. Will it rain today? Will I pass the next math test? Will my favorite team win the toss? Will I get a promotion in next 6 months? All these questions are examples of uncertain situations we live in. Let us map them to few common terminologies which we will use going forward.

* **Experiment –** are the uncertain situations, which could have multiple outcomes. Whether it rains on a daily basis is an experiment.
* **Outcome** is the result of a single trial. So, if it rains today, the outcome of today’s trial from the experiment is “It rained”
* **Event** Each possible outcome of a variable is referred to as an event. An event that has no chance of occurring has a probability of 0. (impossible event.) An event that is sure to occur has a probability of 1. (Certain Event)
* **Probability**is a measure of how likely an event is. So, if it is 60% chance that it will rain tomorrow, the probability of Outcome “it rained” for tomorrow is 0.6

**What is Probability?**

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.

### Why do we need probability?

In an uncertain world, it can be of immense help to know and understand chances of various events. You can plan things accordingly. If it’s likely to rain, I would carry my umbrella. If I am likely to have diabetes on the basis of my food habits, I would get myself tested. If my customer is unlikely to pay me a renewal premium without a reminder, I would remind him about it.

So knowing the likelihood might be very beneficial.

**nPr=**

**Problems**

* 1. Two coins are tossed 500 times, and we get:

Two heads: 105 times

One head: 275 times

No head: 120 times

Find the probability of each event to occur.

**Solution:** Let us say the events of getting two heads, one head and no head by E1, E2 and E3, respectively.

P(E1) = 105/500 = 0.21

P(E2) = 275/500 = 0.55

P(E3) = 120/500 = 0.24

The Sum of probabilities of all elementary events of a random experiment is 1.

P(E1)+P(E2)+P(E3) = 0.21+0.55+0.24 = 1

* 1. The percentage of marks obtained by a student in the monthly tests are given below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test | 1 | 2 | 3 | 4 | 5 |
| Percentage of marks obtained | 69 | 71 | 73 | 68 | 74 |

Based on the above table, find the probability of students getting more than 70% marks in a test.

Solution: The total number of tests conducted is 5.

The number of tests when students obtained more than 70% marks = 3.

So, P(scoring more than 70% marks) = ⅗ = 0.6

**What is the nCr Formula?**

nCr formula is also known as the **"combinations formula".** nCr formula is used to find the number of ways of choosing **r** objects from **n** objects where the order is not important. It is represented in the following way.

**nCr = =**

**Ex1.** Find the values of 10C3,8C4,9P2,5P3

**Solution: R-code with output**

|  |
| --- |
| > c1=choose(10,3); c2=choose(8,4);***#choose() is used to calculate combinations***  > r1=1:2;p1=prod(r1)\*choose(9,2);  > r2=1:3;p2=prod(r2)\*choose(5,3);  > cat("10C3=",c1,"\n");  **10C3= 120**  > cat("8C4=",c2,"\n");  **8C4= 70**  > cat("9C2=",p1,"\n");  9C2= 72  > cat("5P3=",p2,"\n");  **5P3= 60** |

**Ex2.** In how many ways can committee of 4 persons be formed from 4 teachers and 8 students such that exactly 2 teachers will be included in the committee?

**Solution: R-code with output**

|  |
| --- |
| > c1=choose(4,2)\*choose(8,2);  > cat("Total no. of ways is =",c1,"\n");  **Total no. of ways is = 168** |

**Ex3.** What is the probability of drawing 2 ace cards from a well shuffled pack of 52 playing cards?

**Solution: R-code with output**

|  |
| --- |
| > p1=choose(4,2)/choose(52,2);  > cat("Required probability =",p1,"\n");  **Required probability = 0.004524887** |

**Ex4.** A box contains 5 red and 7 blue marbles. A sample of 4 is drawn at random. What is the probability of selecting at least 2 blue marbles?

**Solution: R-code with output**

|  |
| --- |
| > x=0:2;  > p1=sum((choose(5,x)\*choose(7,4-x))/choose(12,4));  > cat("Required probability=",p1,"\n");  **Required probability= 0.8484848** |

**Distribution**

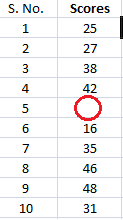
The Possible values a variable can take and how frequently they occur.

## What is Probability Distribution?

A *probability distribution* is a mathematical function that defines the likelihood of different outcomes or values of a variable. This function is commonly represented by a graph or probability table, and it provides the probabilities of various possible results of an experiment or random phenomenon based on the sample space and the probabilities of events. Probability distributions are fundamental in probability theory and statistics for analyzing data and making predictions.

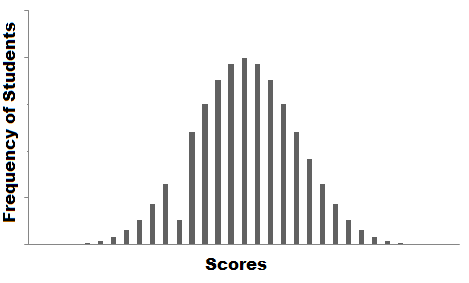
Let’s imagine you’ve collected occupational data for 500 people living in New Delhi. The different possible outcomes are all the various job titles within your dataset. Because occupation is ***categorical in nature (i.e. not numerical*),** the distribution of your dataset would tell you how many (or what percentage) of the people in your sample fall into each group. For example, 20% of the samples are lawyers, 10% are teachers, 5% are nurses, and so on.

Let’s start with an example. Suppose you are a teacher at a university. After checking assignments for a week, you graded all the students. You gave these graded papers to a data entry guy in the university and told him to create a spreadsheet containing the grades of all the students. But the guy only stores the grades and not the corresponding students.



He made another blunder; he missed a few entries in a hurry, and we have no idea whose grades are missing. One way to find this out is by visualizing the grades and seeing if you can find a trend in the data.

With ***numerical data***, the distribution will order the data from lowest to highest value. In this case, the distribution is presented as a ***graph or chart***. The trained eye can then look at the shape of the graph to see, at a glance, how the data is distributed. A so-called normal distribution produces a symmetrical, bell-shaped curve on a graph.



The graph you plotted is called the [frequency distribution](http://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=6ProbabilityDistributionsarticle) of the data. You see that there is a smooth curve-like structure that defines our data, but do you notice an anomaly? We have an abnormally low frequency at a particular score range. So the best guess would be to have missing values that remove the dent in the distribution.

Based on the type of numeric data, the probability distribution is classified into two types.

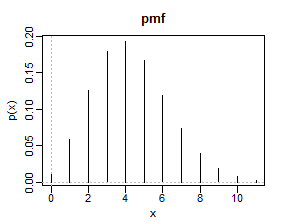
1. **Discrete Probability Distribution**: For a finite number of outcomes- Ex: Roll dice, Picking a Card from a pack of 52 cards.
   1. Bernoulli Distribution
   2. Binomial Distribution
   3. Poisson Distribution
2. **Continuous Probability Distribution:** for infinitely many outcomes- Ex: Recording time, Distance in tracking field
   1. Normal
   2. Uniform
   3. Exponential

In probability and statistics, several terms are used to describe the various functions that are used to model probability distributions. These include:

1. **Probability mass functions**

The PMF is a function that describes the probability of a **discrete** random variable **taking on a certain value.** It is a mathematical function that describes the probability that a random variable will take on a specific value rather than falling within a range of values.

**f(x)=P(X=x)**

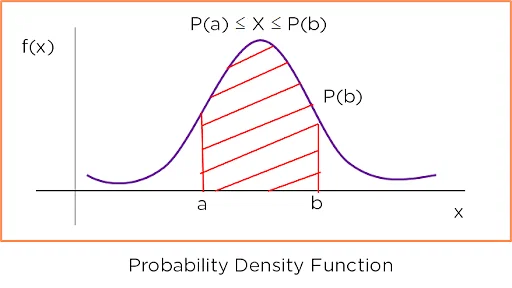


1. **Probability density functions**

A function that defines the relationship between a random variable and its probability, such that you can find the probability of the variable using the function, is called a Probability Density Function (PDF) in statistics.

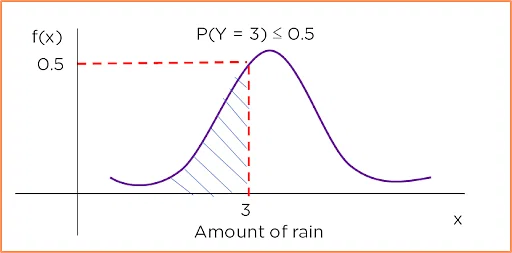
The PDF is a function that describes the probability of a **continuous** random variable **taking on a certain value.** It is a mathematical function that describes the probability that a random variable will fall within a certain range of values.

**P(a) <= X <= P(b)**



*Fig. Probability Density Function*

Consider the graph below, which shows the rainfall distribution in a year in a city. The x-axis has the rainfall in inches, and the y-axis has the probability density function. The probability of some amount of rainfall is obtained by finding the area of the curve on the left of it.



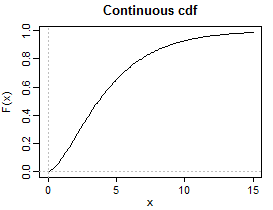
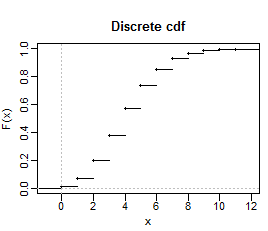
For the probability of 3 inches of rainfall, you plot a line that intersects the y-axis at the same point on the graph as a line extending from 3 on the x-axis does. This tells you that the probability of 3 inches of rainfall is less than or equal to 0.5.

1. **Cumulative distribution function (CDF):**

The CDF is a function that describes the probability that a random variable **(continuous or discrete)** will **take on a value less than or equal to a certain value**. It is obtained by summing up the probability density function and getting the cumulative probability for a random variable. The cumulative distribution function of a random variable to be calculated at a point x is represented as **Fx(X).** It is the probability that the random variable X will take a value less than or equal to x.

The formula depicted below shows the cumulative distribution function calculated between points (a, b) for the PDF Fx(x).

**Fx(X)=P(X<=x)**

*Fig. Cumulative Distributive Function*

* 1. **Bernoulli Distribution**

Bernoulli distribution is a discrete probability distribution, meaning it’s concerned with **discrete random variables**. A discrete random variable is one that has a finite or countable number of possible values—the number of heads you get when tossing three coins at once, or the number of students in a class.

So: A discrete probability distribution describes the probability that each possible value of a discrete random variable will occur—for example, the probability of getting a six when rolling a die. When dealing with discrete variables, the probability of each value falls between 0 and 1, and the sum of all the probabilities is equal to 1. So, in the die example, assuming we’re using a standard die, the probability of rolling a six is 0.167, or 16.7%. This is based on dividing 1 (the sum of all probabilities) by 6 (the number of possible outcomes).

**Bernoulli distribution and Bernoulli trials explained**

**Bernoulli event:** An event for which the probability of occurrence is p and the probability of the event not occurring is 1-p i.e., the event has only two possible outcomes (these can be viewed as Success or Failure, Yes or No and Heads or Tails). The event occurs with a probability p and 1-p respectively.

**Bernoulli’s Trials**

Bernoulli’s trials are a type of experiment where you repeat the same thing over and over again independently (That means the probabilities must remain the same throughout the trials; each event must be completely separate and have nothing to do with the previous event). Each time, you’re looking for a specific outcome, which we call **event A**. The chance of this outcome happening is the same each time you do the experiment, no matter how many times you repeat it.

The experiment itself only has two possible outcomes, either event A happens, or it doesn’t. It’s like flipping a coin where you can either get heads or tails.

**Examples of Bernoulli trials:**

* If an item produced in a factory is selected at random, it may be either defective or non-defective. Such a trial which results in only two possible outcomes i.e, a success and other a failure.
* **Coin tosses:** Record how many tosses of coins resulted in heads and how many coin tosses resulted in tails. We can consider the result of getting heads as success and not getting head i.e., getting tails to be a failure.
* **Football:** How many shots on a goal post resulted in the goal score, and how many shots were missed. We can call a goal scored as a “success” and a missed target to be a failure.
* **Rolling Dice:** The probability of a roll of two dice resulting in a double six. A double six dice roll could be considered to be a success and everything else can be considered a failure.
* **Births:** how many boys are born and how many girls are born each day.

**Bernoulli process:** A sequence of Bernoulli trials is called a Bernoulli process. Among other conclusions that could be reached, for n trials, the probability of n successes is pⁿ.

**Bernoulli distribution**

Bernoulli distribution applies to events that have one trial and two possible outcomes. It represents the success or failure of a single Bernoulli trial. Think of any kind of experiment that asks a yes or no question—for example, will this coin land on heads when I flip it? Will I roll a six with this die? Will I pick an ace from this deck of cards? Will voter X vote “yes” in a political referendum? Will student Y pass their math test?

**Bernoulli distribution example: Tossing a coin**

The coin toss example is perhaps the easiest way to explain Bernoulli distribution. Let’s say that the outcome of “heads” is a “success,” while an outcome of “tails” is a “failure.” In this instance:

The probability of a ***successful outcome*** (landing on heads) is written as **p**

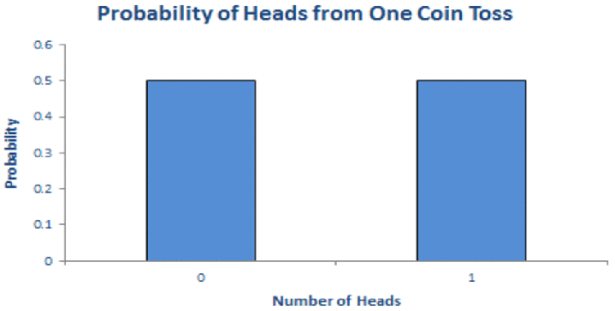
The probability of a ***failure*** (landing on tails), written as q, is calculated as **1 - p**

With a standard coin, we know that there’s a 50/50 chance of landing on either heads or tails. So, in this case:

p = 0.5

q = 1–0.5

So, in our coin toss example, both p and q = 0.5. On a graph, you’d represent the probability of a failure as “0” and the probability of success as “1,” both on the y-axis.



**Properties of a Bernoulli distribution:**

* There are only two possible outcomes a 1 or 0, i.e., success or failure in each trial.
* The probability values of mutually exclusive events that encompass all the possible outcomes need to sum up to one.
* If the probability of success is p then the probability of failure is given as 1-p. The probability values must remain the same across each successive trial. Each event must be completely separate and have nothing to do with the previous event. i.e., the probabilities are not affected by the outcomes of other trials which means the trials are independent.

The **expected value** for a random variable, X, from a Bernoulli distribution can be given as-

**E[X] = 1\*(p) +0\*(1-p) = p, for example if p=0.6, then E[X] =0.6**

The **mean** of Bernoulli random variable(X) is

**E[X] = 1(p) +0(1-p)**

**= p**

The **variance** of Bernoulli random variable is

**V[X] = E[X²]-[E(X)]² = 1²p+0²(1-p)-p²**

**= p(1-p)**

**Applications of Bernoulli Outcomes**

There are real-life situations that involve noting if a specific event occurs or not. Such events are recorded as a success or a failure. E.g. Some of the examples that explain binary outcome scenarios involve calculating the probability of-

* Success of a medical treatment
* Interviewed person being a female
* Student result(Pass/fail) in an exam
* Transmittance of a disease (transmitted/not transmitted)
  1. **Binomial Distribution**

The Binomial Distribution represents the number of successes and failures in n independent Bernoulli trials for some given value of n. **Binomial distribution** *is the discrete probability distribution of the number of success in a sequence of n independent Bernoulli trials (having only yes/no or true/false outcomes).*

**Example:**

Let’s say your football team is playing a series of 5 games against your opponent. Who ever wins more games (out of 5) wins the title.

Let us say, your team might is more skilled and has 75% chances of winning. So, there is a 25% chance of losing it.

What is the probability of you winning the series? Is it 75% or is it something else?

Let us find out. What are the possible scenarios in playing 5 games?

WWWWW, WWWWL, WWWLL, WWLLL, WLLLL, LLLLL, LWWWW and so on….

So for the first game, there are two possibilities, you either win or lose, again for the second game we have two possibilities. Assuming that the first game has no effect on the outcome of the second – No one gets tired, no one gets under pressure after losing etc.

So let’s define our random variable X to be a number of wins in 5 games. Remember probability of winning is 0.75 and losing is 0.25. Assume that a tie doesn’t happen.

X=Number of wins in 5 games

So the first game has 2 outcomes – win and lose, second again has 2 and so on.

So total possibilities is 2\*2\*2\*2\*2 = 32

P (X=0) denotes the probability that you lose all the games and there is only one way that can happen i.e. {LLLLL} = 0.25\*0.25\*0.25\*0.25\*0.25 (multiplying the probabilities of losing the each time, lost first time and second time and third time and so on..)

P(X=1) denotes the probability that you win only 1 game i.e.(WLLLL or LWLLL or LLWLL or LLLWL or LLLLW). So there are 5 cases where you win 1 game = 5\*0.75\*0.25\*0.25\*0.25\*0.25=0.0146

While we can count each of these possible outcomes, it becomes very exhaustive and intensive exercise. Let us take help of combinatorics here. Choose 2 wins out of 5 games = 5C2 ()

so, the Probability for getting k successes in n Bernoulli trails is given by:

P(X=k) = nCk pk qn-k  ,  [here p is the probability of success and q is the probability of failure]

Let’s see how this comes.

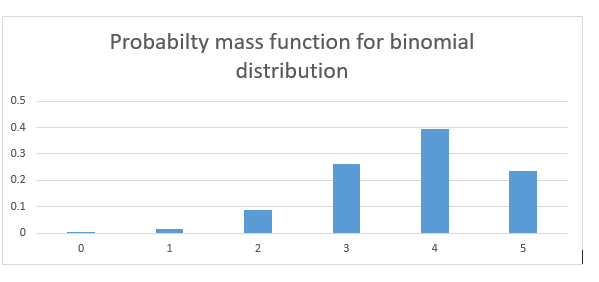
P(X=2) denotes the probability that you win 2 games. So there are 5C2() = 10 cases where you win 2 games. Hence probability = 10\*0.75\*0.75\*0.25\*0.25\*0.25=0.088

P(X=3) denotes the probability that you win 3 games. So, there are 5C3() =10 cases where you win 3 games. Hence probability = 10\*0.75\*0.75\*0.75\*0.25\*0.25=0.264

Similarly,  P(X=4) = 0.395

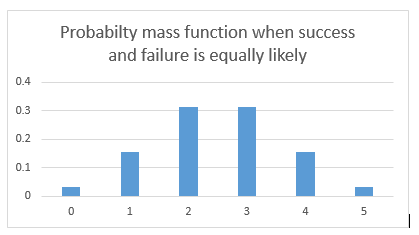
P(X=5) = 0.237

* What we just calculated were discrete probabilities for a Binomial distribution. If we look at these probabilities we get something like:



**As you can see the probability of winning the series is much higher than 0.75.**

* If the events are equally likely to occur i.e. p = q = 0.5, the probability distribution looks something like the graph below. Here the probability of success and failure is the same.



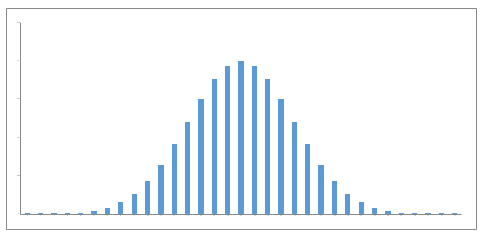
What difference do we see in the two probability distributions?  The first one is skewed towards right. Reason being the likelihood to win is more, hence more wins are more likely than more losses.

In the second case when wins and losses are equally likely, so the distribution is symmetrical.

 Let’s assume that probability of winning and losing is equal. p=q=0.5

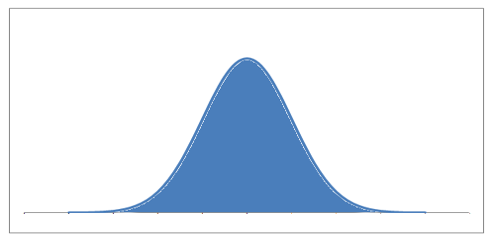
* Now, What if I increase my number of trials? What if I play 20 games of football with a probability of winning and losing to be 50-50? There are a lot more possibilities and combinations. The bars get thinner and thinner.

The bars get thinner and thinner.



* What if I play an infinite number of times with equal probability for winning and losing?

The bars get infinitely small and the probability distribution looks something like a continuous set of bars which are very close, almost continuous. This now becomes a probability density function. Notice that this now becomes a continuous function.



The main relation between Bernoulli distribution and binomial distribution arises with the number of times a trial is performed. The binomial distribution tries to summarize the number of successes k in a given number of Bernoulli trials n, with a probability of success for each trial.

The mathematical representation of binomial distribution is given by:

**P(x)=** Px qn-x

where,

* p is the Probability of Success,
* q = 1 – p is the Probability of Failure,
* n is the Number of Independent trials, and
* x is the number of times an event occurred.

The mean and variance of a binomial distribution are given by:

**Mean** -> µ = n\*p

**Variance** -> Var(X) = n\*p\*q

Here are some real-life scenarios where the binomial distribution is applicable:

1. **Coin Tossing:** Suppose you toss a fair coin ten times. The binomial distribution can be used to determine the probability of getting a certain number of heads (or tails). Each coin toss is an independent event, the probability of getting a head or tail is the same for each toss (0.5), and you are conducting a fixed number of trials (10 coin tosses).
2. **Quality Control in Manufacturing:** A factory produces items and each item may be defective or not defective. If you randomly select a certain number of items (fixed number of trials), the binomial distribution can be used to calculate the probability of finding a specific number of defective items. Each selection is an independent event and the probability of selecting a defective item is the same for each selection.
3. **Medical Trials:** Suppose a drug has a 70% chance of curing a certain disease. If the drug is given to 50 patients (fixed number of trials), the binomial distribution can help estimate the probability of the drug curing a certain number of patients. Each patient's outcome is independent of the others and the probability of success (curing the disease) is the same for each trial.
4. **Survey Sampling:** If you're conducting a survey and you know that 60% of a population will choose option A (based on past data or a larger sample), a binomial distribution can help you determine the probability that a certain number out of a smaller sample will choose option A. Each survey response is an independent event, the probability of each person choosing option A is the same, and you're surveying a fixed number of people.
5. **Sports:** If a basketball player makes a free throw 80% of the time, you can use the binomial distribution to calculate the probability of that player making a certain number of free throws out of a fixed number of attempts. Each free throw is an independent event, and the probability of success is the same for each shot.

The difference between Bernoulli distribution and binomial distribution is given below:

|  |  |
| --- | --- |
| **Bernoulli Distribution** | **Binomial Distribtuion** |
| Bernoulli distribution is used when we want to model the outcome of a single trial of an event. | If we want to model the outcome of multiple trials of an event, Binomial distribution is used. |
| It is represented as X ∼∼ Bernoulli (p). Here, p is the probability of success. | It is denoted as X ∼∼ Binomial (n, p). Where n is the number of trials. |
| Mean, **E[X] = p** | Mean, **E[X] = np** |
| Variance, **Var[X] = p(1-p)** | Variance, **Var[X]= np(1-p)** |
| **Example:** Suppose the probability of passing an exam is 80% and failing is 20%. Then the Bernoulli distribution can be used to model the passing or failing in such an exam. | **Example:** Suppose the probability of passing an exam is 80% and failing is 20%. Then if we want to find the probability that a student will pass in exactly 4 out of 5 exams, we use the Binomial Distribution. |

R supports following functions related to binomial distribution with specified parameters

|  |  |  |
| --- | --- | --- |
| Functions | **Arguments** | **Description** |
| **dbinom(**x, size, prob) |  | It allows users to calculate the probability of obtaining a specific number of "successes" in a fixed number of Bernoulli trials, given a certain probability of success. |
| **pbinom(**q,size,prob**)** | The **'q'** parameter is the number of successes we're interested in, **'size'** represents the number of trials, and **'prob'** is the probability of success on each trial. | This function is incredibly helpful when we need to compute the probability of having a certain number of successes or fewer in a given number of independent trials. |
| **qbinom(**p,size,prob**)** | **'p'** is the percentile we're interested in (expressed as a probability), **'size**' is the number of trials, and **'prob'** is the probability of success on each trial. | It is used for a binomial distribution to find the quantile function, or the number of successes at a given percentile. |
| **rbinom(**n,size,prob**)** | **'n'** is the number of random numbers we want to generate, **'size'** is the number of trials, and **'prob'** is the probability of success on each trial. | It allows us to generate random numbers following a binomial distribution. |

**Ex1.** Here we're looking at a situation with 5 trials (let's say flipping a coin 5 times), and we want to know the probability of getting 3 successes (let's say 3 heads). Assuming a fair coin, the probability of success on each trial is 0.5.

**Solution: R-Code with output**

|  |
| --- |
| > prob\_three\_heads <- dbinom(3, size=5, prob=0.5)  > print(prob\_three\_heads)  **[1] 0.3125** |

**Ex. 2** Let's say we're interested in the probability of seeing 5 or fewer successes in 20 trials, with the probability of success on each trial being 0.25.

**Solution: R-Code with output**

|  |
| --- |
| > prob\_five\_or\_less <- pbinom(5, size=20, prob=0.25)  > print(prob\_five\_or\_less)  **[1] 0.6171727** |

This indicates that the probability of getting 5 or fewer successes in 20 trials, with the probability of success on each trial being 0.25, is approximately 0.6171.

**Ex3.** We have 10 trials (e.g., flipping a coin 10 times), and we want to know the cumulative probability of getting 4 successes (e.g., 4 heads) or less. Assuming a fair coin, the probability of success on each trial is 0.5.

**Solution: R-Code with output**

|  |
| --- |
| > prob\_four\_or\_less <- pbinom(4, size=10, prob=0.5)  > print(prob\_four\_or\_less)  **[1] 0.3769531** |

**Ex4.** Let's say we have 10 trials (e.g., flipping a coin 10 times) and we want to know the number of successes (e.g., heads) we'd expect to see at the 70th percentile. Assuming a fair coin, the probability of success on each trial is 0.5.

**Solution: R-Code with output**

|  |
| --- |
| > num\_successes <- qbinom(0.7, size=10, prob=0.5)  > print(num\_successes)  **[1] 6** |

**Ex5.** We want to generate 100 random numbers (let's say 100 experiments of flipping a coin 10 times), and we're interested in each experiment's number of successes (e.g., heads). Assuming a fair coin, the probability of success on each trial is 0.5.

**Solution: R-Code with output**

|  |
| --- |
| > random\_numbers <- rbinom(100, size=10, prob=0.5)  > print(random\_numbers)  **[1] 4 5 3 4 3 4 4 4 5 5 3 3 4 4 7 5 8 6 6 5 4 6 5 5 6 4 5 3 8 5 7 3 6 5 4 6 5**  **[38] 6 3 4 6 7 4 3 5 5 5 5 5 2 4 5 3 5 4 4 4 7 7 4 3 3 5 4 6 7 4 5 4 4 5 6 5 7**  **[75] 6 7 4 6 6 9 4 6 7 8 7 4 3 5 3 6 4 4 6 5 4 7 3 7 6 3** |

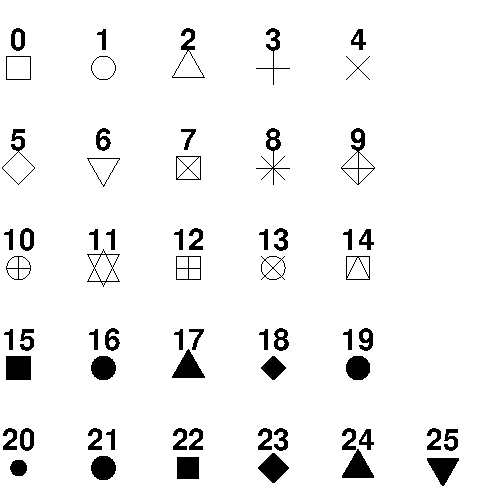
These are the number of successes in each of the 100 experiments. For instance, in the first experiment, we got 4 heads, in the second experiment; we got 5 heads, and so on.

**Ex6.** Draw a random sample of size 8 from a Binomial Distribution with n=6 and p=1/3. Find the mean of sample

**Solution: R-Code with output**

|  |
| --- |
| > r=rbinom(8,6,1/3);  > m=mean(r);  > cat("Required sample is :",r,"\n");  **Required sample is : 0 3 1 1 3 1 2 1**  > cat("Mean of sample=",m,"\n");  **Mean of sample= 1.5** |

**Note :** pch🡪 point shape ranges from 0 to 25



The graph of probability mass function of binomial distribution for different values of parameters n and p are given below.

* 1. n=10, p=0.4

|  |
| --- |
| > n=10; p=0.4;x=0:n;  > px=dbinom(x,n,p);  > plot(x,px,"h",main="Binomial probability distribution",xlab="X",ylab="p(x)");  > points(x,px,pch=16);  > legend("topleft",legend = c("Binomial(n=10,p=0.4"),pch = 16) |

* 1. n=5, p=0.6

|  |
| --- |
|  |

* 1. n=18, p=0.6

|  |
| --- |
|  |

In fig, we observe that as n increases binomial distribution approaches to normal.

**Ex.7** The frequency distribution of number of heads obtained in an experiment of tossing 5 coins 110 times is given below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No. of heads** | 0 | 1 | 2 | 3 | 4 | 5 |
| **Frequency** | 6 | 15 | 25 | 42 | 18 | 4 |

Fit a binomial distribution to the above data and find the expected frequencies. Plot observed and expected frequencies and comment on the adequacy of model.

|  |
| --- |
| > x=0:5;  > f=c(6,15,25,42,18,4);  > m=sum(f\*x)/sum(f);  > n=max(x);  > p=m/n;  > q=1-p;  > px=dbinom(x,n,p);  > px1=round(px,6);  > ef=sum(f)\*px;  > ef1=round(ef,0);  > d=data.frame(x,f,"fx"=f\*x,"Probability"=px1,"Exp.Freq"=ef1);  > print(d)  **x f fx Probability Exp.Freq**  **1 0 6 0 0.026961 3**  **2 1 15 15 0.142885 16**  **3 2 25 50 0.302896 33**  **4 3 42 126 0.321047 35**  **5 4 18 72 0.170143 19**  **6 5 4 20 0.036068 4**  > cat("Summ f=",sum(f),"\n");  **Summ f= 110**  > cat("Sum f\*x=",sum(f\*x),"\n");  **Sum f\*x= 283**  > cat("Mean frequency distribution=",m,"\n");  **Mean frequency distribution= 2.572727**  > plot(f,ef1,xlab="Observed Frequency",ylab="Expected Frequency","p",pch=16);  > abline(0,1); |

* 1. **Poisson Distribution**

A Poisson distribution is a discrete probability distribution, meaning that it gives the probability of a discrete (i.e., countable) outcome. For Poisson distributions, the discrete outcome is the number of times an event occurs, represented by k.

You can use a Poisson distribution to predict or explain the number of events occurring within a given interval of time or space. “Events” could be anything from disease cases to customer purchases to meteor strikes. The interval can be any specific amount of time or space, such as 10 days or 5 square inches.

**You can use a Poisson distribution if:**

* Individual events happen at random and independently. That is, the probability of one event doesn’t affect the probability of another event.
* You know the mean number of events occurring within a given interval of time or space. This number is called λ (lambda), and it is assumed to be constant.

**Examples of Poisson distributions**

A Poisson distribution could be used to explain or predict:

* Text messages per hour
* Male grizzly bears per hectare
* Machine malfunctions per year
* Website visitors per month
* Influenza cases per year

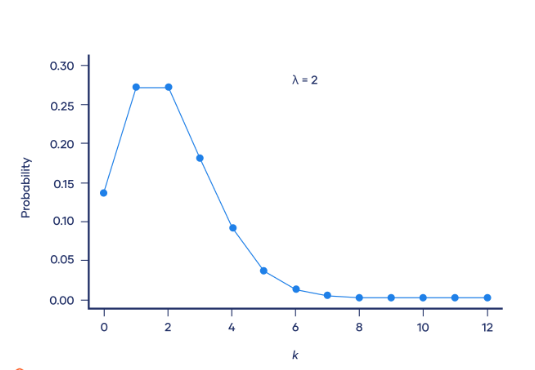
**Probability mass function graphs**

A Poisson distribution can be represented visually as a graph of the probability mass function. A probability mass function is a function that describes a discrete probability distribution.

The most probable number of events is represented by the peak of the distribution—the **mode**.

* When λ is a non-integer, the mode is the closest integer smaller than λ.
* When λ is an integer, there are two modes: λ and  λ−1.

When λ is low, the distribution is much longer on the right side of its peak than its left (i.e., it is strongly **right-skewed**).



**Mean and variance of a Poisson distribution**

The Poisson distribution has only one parameter, called λ.

* The **mean**of a Poisson distribution is λ.
* The **variance** of a Poisson distribution is also λ.

In most distributions, the mean is represented by µ (mu) and the variance is represented by σ² (sigma squared). Because these two parameters are the same in a Poisson distribution, we use the λ symbol to represent both.

Poisson distribution formula

The probability mass function of the Poisson distribution is:

P(X=x)=

Where:

X is a random variable following a Poisson distribution

x is the number of times an event occurs

P(X=x) is the probability that an event will occur x times

e is Euler’s constant (approximately 2.718)

λ is the average number of times an event occurs

! is the factorial function

**Example: Applying the Poisson distribution formula**

An average of 0.61 soldiers died by horse kicks per year in each Prussian army corps. You want to calculate the probability that exactly two soldiers died in the VII Army Corps in 1898, assuming that the number of horse kick deaths per year follows a Poisson distribution.

**Calculation**

The specific army corps (VII Army Corps) and year (1898) don’t matter because the probability is constant.

 x = 2 deaths by horse kick

 λ = 0.61 deaths by horse kick per year

 e = 2.718

P(X=k)=

P(X=2)=

P(X=2)=

P(X=2)=0.101

**Assignment 1**

At a small walk-in clinic, an average of five patients arrives at the clinic per hour during opening hours. What is the probability that exactly three patients will arrive in the next hour?

**The Poisson distribution function used in R**

|  |  |
| --- | --- |
| **Function** | **Description** |
| **dpois** | Poisson probability mass function (Probability function) |
| **ppois** | Poisson distribution (Cumulative distribution function) |
| **qpois** | Poisson quantile function |
| **rpois** | Poisson pseudorandom number generation |

**Ex1.** Implementing the example problem in R

|  |
| --- |
| > m=0.61;x=2;px=dpois(x,m);  > px  **[1] 0.1010904** |

**Ex2.** The graph of Poisson probability mass function for m=4

**Solution: R-code with output**

|  |
| --- |
| > m=4;x=0:25;px=dpois(x,m)  > plot(x,px,"h",main="Poisson probability curve", xlab="X", ylab="Probability")  > points(x,px,pch=16)  > legend("topright",legend = c("Poisson (m=4,x=25"),pch = 16) |

**Ex3.** The behaviour of Poisson distribution for large values of m

|  |  |
| --- | --- |
|  |  |
|  |  |

From the above graph, we observe that as the value of m increases the Poisson distribution tends to normal disttribution.

**Ex4.** Draw a random sample of size 8 from a poisson distribution with parameter 1.7

**Solution: R-code with output**

|  |
| --- |
| > x=8;m=1.7;  > s=rpois(x,m)  > cat("Required sample is:",s,"\n");  **Required sample is: 0 1 3 0 4 0 2 0** |

**Ex5.**Draw a random sample of size 5 from a Poisson distribution with mean=0.8. Find the mean of your sample.

**Solution: R-code with output**

|  |
| --- |
| > s=rpois(5,0.8)  > ms=mean(s)  > cat("Required sample is :",s,"\n");  **Required sample is : 2 1 0 1 2**  > cat("Mean of the sample is:",ms,"\n");  **Mean of the sample is: 1.2** |

**Ex6.** A car hire firm has 2 cars. The number of demands for a car on a day has Poisson distribution with mean 1.5. Find the probability that on a day (i) neither car is used (ii) Some demand is not fulfilled.

**Solution: R-code with output**

|  |
| --- |
| > m=1.5;x=2;  > rp1=dpois(0,m)  > rp2=1-ppois(2,m)  > cat("Prob. that neither car is used is:",rp1,"\n")  **Prob. that neither car is used is: 0.2231302**  > cat("Prob. that some demand is not fulfilled:",rp2,"\n")  **Prob. that some demand is not fulfilled: 0.1911532** |

**Ex7.** If the probability that an individual suffers from a bad reaction from injection of a serum is 0.001, determine the probability that out of 2000 individuals injected 2 or more will suffer from a bad reaction.

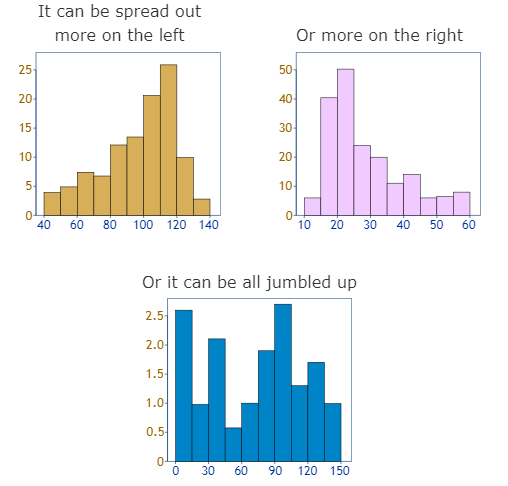
**Solution: R-code with output**

|  |
| --- |
| > n=2000;p=0.001;  > m=n\*p;  > rp=1-ppois(1,m)  > cat("Required probability is=",rp,"\n");  **Required probability is= 0.5939942** |

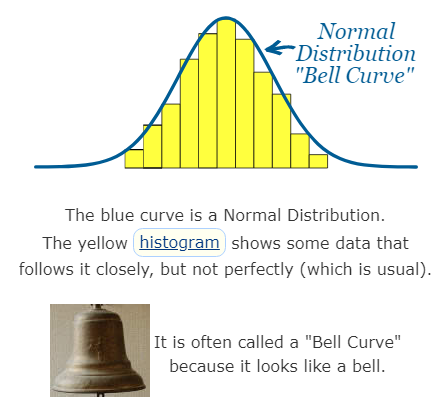
**2.1 Normal Distribution**

The Normal Distribution, also called the Gaussian distribution, is the most significant continuous probability distribution.

Data can be "distributed" (spread out) in different ways.



But there are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" like this:



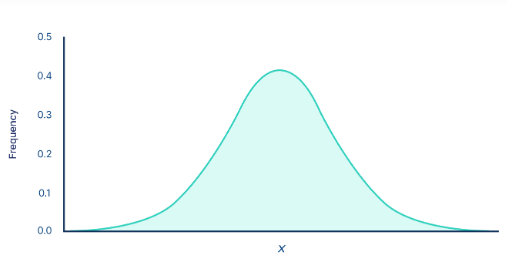
Many things closely follow a Normal Distribution:

* heights of people
* size of things produced by machines
* errors in measurements
* blood pressure
* marks on a test

**What are the properties of normal distributions?**

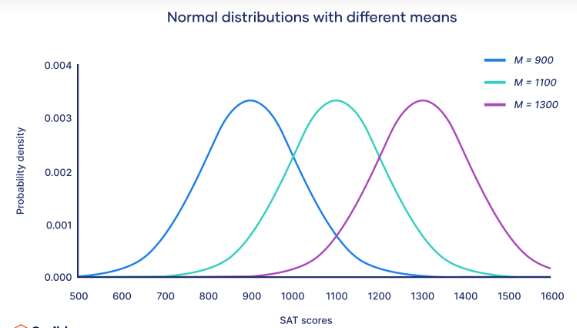
Normal distributions have key characteristics that are easy to spot in graphs:

* The **mean, median** and **mode** are exactly the same.
* The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
* The distribution can be described by two values: the **mean** and the **standard deviation**.

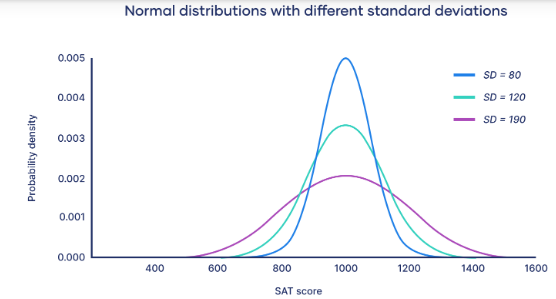
****

The **mean** is the location parameter while the standard deviation is the scale parameter.

The mean determines where the peak of the curve is centered. Increasing the mean moves the curve right, while decreasing it moves the curve left.



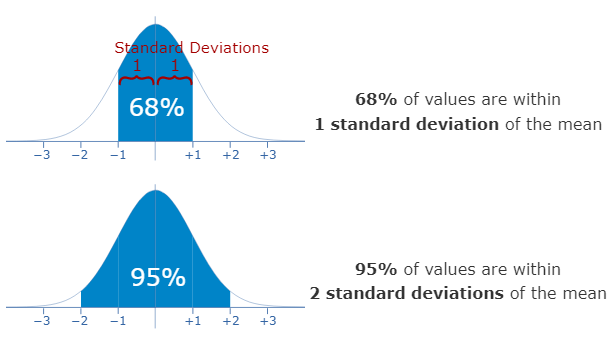
The **standard deviation** stretches or squeezes the curve. A small standard deviation results in a narrow curve, while a large standard deviation leads to a wide curve.

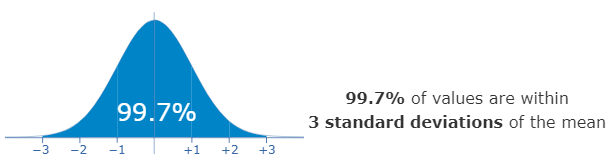
****

**Empirical rule**

The empirical rule, or the 68-95-99.7 rule, tells you where most of your values lie in a normal distribution:

When we calculate the standard deviation we find that generally:



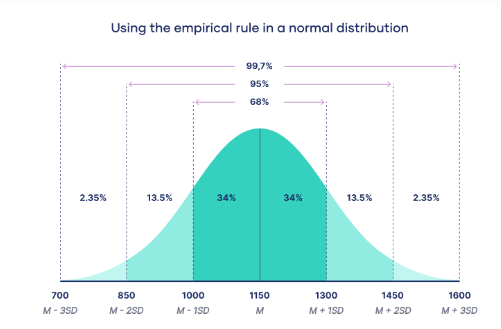


**Example 1: Using the empirical rule in a normal distribution**

You collect SAT scores from students in a new test preparation course. The data follows a normal distribution with a mean score (M) of 1150 and a standard deviation (SD) of 150.

**Following the empirical rule:**

* Around 68% of scores are between 1,000 and 1,300, 1 standard deviation above and below the mean.
* Around 95% of scores are between 850 and 1,450, 2 standard deviations above and below the mean.
* Around 99.7% of scores are between 700 and 1,600, 3 standard deviations above and below the mean.



The empirical rule is a quick way to get an overview of your data and check for any outliers or extreme values that don’t follow this pattern.

**Example 2: 95% of students at school are between 1.1m and 1.7m tall.**

Assuming this data is normally distributed can you calculate the mean and standard deviation?

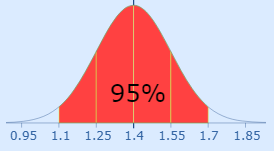
The mean is halfway between 1.1m and 1.7m:

Mean = (1.1m + 1.7m) / 2 = 1.4m

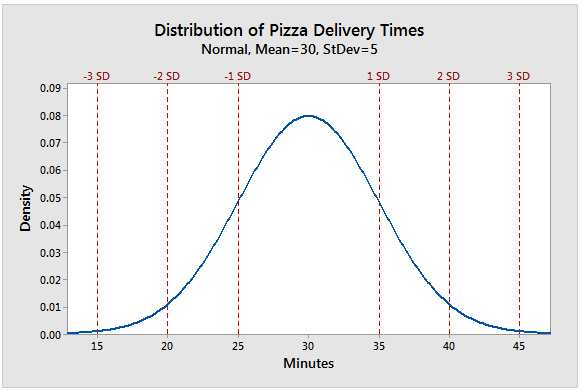
95% is 2 standard deviations either side of the mean (a total of 4 standard deviations) so:

|  |  |
| --- | --- |
| 1 standard deviation | = (1.7m-1.1m) / 4 |
|  | = 0.6m / 4 |
|  | = 0.15m |

And this is the result:

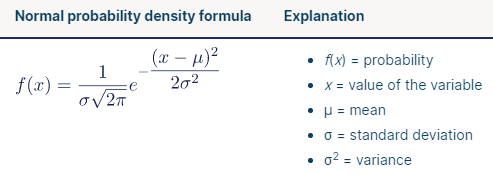


**Example 3:** Let’s look at a pizza delivery example. Assume that a pizza restaurant has a mean delivery time of 30 minutes and a standard deviation of 5 minutes. Using the Empirical Rule, we can determine that 68% of the delivery times are between 25-35 minutes (30 +/- 5), 95% are between 20-40 minutes (30 +/- 2\*5), and 99.7% are between 15-45 minutes (30 +/-3\*5). The chart below illustrates this property graphically.



**Formula of the normal curve**

Once you have the mean and standard deviation of a normal distribution, you can fit a normal curve to your data using a **probability density function**.



**Example 1**. Calculate the probability of normal distribution with the population mean 2, standard deviation 3 or random variable 5.

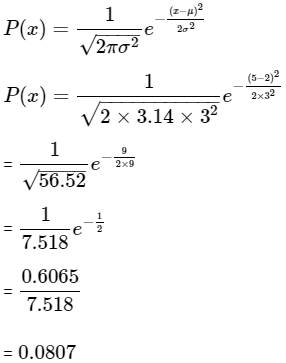
**Solution:**

x = 5

Mean = μ = 2

Standard Deviation = σ = 3

We will solve the questions with the help of the above normal probability  distribution formula:



2. An average electric bulb lasts for 300 days with a standard deviation of 50 days. Assume that bulb life is normally distributed, what is the probability that the electric bulb will last at most 365 days.

**Solution :**

Given, Mean score = 300 days

Standard deviation = 50 days

We need to determine the cumulative probability that bulb life is less than or equals to 365 days.

We have the following information:

The normal random variable value is given as 365 days.

The mean is equivalent to 300 days.

The standard deviation is equivalent to 50 days.

Hence, the answer of the question is P (x < 365) = 0.90.

It states that there is a 90% probability that an electric bulb will burn out within 365 days.

**Normal distribution functions**

|  |  |  |
| --- | --- | --- |
|  | **Purpose** | **Syntax** |
| **rnorm** | Generates random numbers from normal distribution | rnorm(n, mean, sd) |
| **dnorm** | Probability Density Function (PDF) | dnorm(x, mean, sd) |
| **pnorm** | Cumulative Distribution Function (CDF) | pnorm(q, mean, sd) |
| **qnorm** | Quantile Function – inverse of pnorm | qnorm(p, mean, sd) |

**Ex1.** The graph of normal probability curve for µ=20, σ=4 can be drawn using following commands

**Solution: R-Code with output**

|  |
| --- |
| > m=20;sd=4  > x=seq(m-4\*sd,m+4\*sd,0.1);  > pr=dnorm(x,m,sd);  > s=c("mu=20","var=16");  > plot(x,pr,"l",col=4,main="Normal probability curve",ylab="f(x)",xlab="x",cex.main=0.8);  > legend("topright",legend = s,horiz=F); |

**Ex2.** Draw a random sample of size 20 from Normal distribution with mean 4 and variance 2. Find mean, median and standard deviation of the sample

**Solution: R-Code with output**

|  |
| --- |
| > s=rnorm(20,4,2^0.5);  > n=20;  > m=mean(s);me=median(s);  > v=((n-1)/n)\*var(s);  > sd=v^0.5;  > cat("sample is :",s,"\n");  **sample is : 2.039117 0.9366973 3.791393 5.040186 5.607668 4.52771 6.373755 3.073111 3.507609 4.962977 2.582421 5.501635 5.90986 5.812284 3.228858 5.756019 4.273225 3.053708 1.529962 5.262312**  > cat("Mean of the sample is:",m,"\n")  **Mean of the sample is: 4.138525**  > cat("Median of the sample is:",me,"\n")  **Median of the sample is: 4.400467**  > cat("Standard Deviation of the sample is:",sd,"\n")  **Standard Deviation of the sample is: 1.549783** |

**Ex3.** The distribution of income of a large group of workers is normal with mean Rs.2000 and standard deviation Rs.200. What proportion of workers secured income (i) between Rs.2400 and Rs. 2500 (ii) more than Rs. 2500?

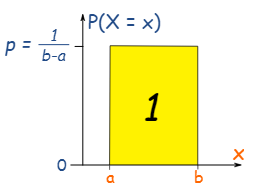
**Solution: R-Code with output**

|  |
| --- |
| > mu=2000;sd=200;  > rp1=pnorm(2500,mu,sd)-pnorm(2400,mu,sd);  > rp2=1-pnorm(2500,mu,sd);  > cat("Proportion of workers having income between 2400 and 2500 is=",rp2,"\n");  **Proportion of workers having income between 2400 and 2500 is= 0.006209665**  > cat("Proportion of workers having income between >2500 is=",rp1,"\n");  **Proportion of workers having income between >2500 is= 0.01654047** |

* 1. **Uniform Distribution**

The Uniform Distribution (also called the Rectangular Distribution) is the simplest distribution.

It has equal probability for all values of the Random variable between a and b:



The probability of any value between a and b is p

We also know that p = 1/(b-a), because the total of all probabilities must be 1, so

🡪The area of the rectangle = 1

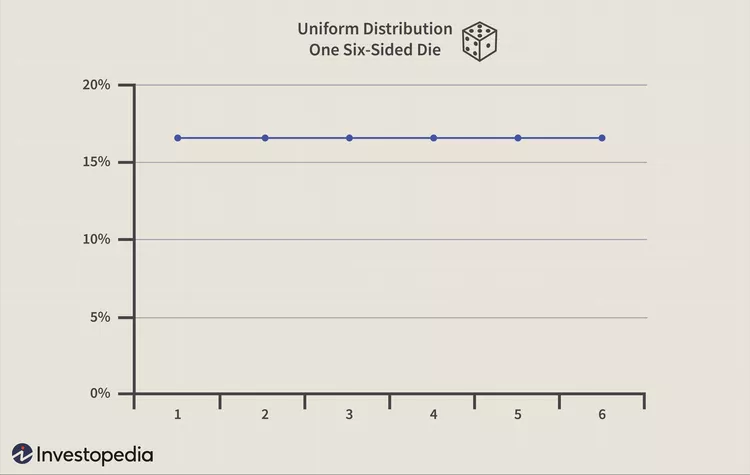
🡪p × (b−a) = 1

🡪p = 1/(b−a)

The uniform distribution is a type of probability distribution in that all the possible outcomes are equally possible. A deck of cards has uniform distributions within it since the probability of drawing a heart, club, diamond or spade is equally possible. A coin also has a uniform distribution since the probability of getting either the heads or the tails in the coin toss is the same.

The uniform distribution can be visualized as the straight horizontal line, hence, for a coin flip returning to a head or a tail, both have a probability p = 0.50 and it would be depicted by the line from the y-axis at 0.50.

The roll of a single dice yields one of six numbers: 1, 2, 3, 4, 5, or 6. because there are only 6 possible outcomes, the probability of you landing on any one of them is 16.67% (1/6). When plotted on a graph, the distribution is represented as a horizontal line, with each possible outcome captured on the x-axis, at the fixed point of probability along the y-axis.



There are two kinds of uniform distributions namely *discrete and continuous*. In the former type of distribution, each of the possible outcomes is discrete. In continuous distribution, the outcomes are continuous and infinite.

**Uniform Probability Distribution**

A continuous probability distribution is called the uniform distribution and it is related to the events that are equally possible to occur. It is defined by two different parameters, x and y, where x = the minimum value and y = the maximum value. It is generally represented by u(x,y).

If the probability density function or the probability distribution of the uniform distribution with a continuous random variable X is  it is denoted by U(x, y) where x and y are the constants in a way that x < a < y. It is written in the following manner: X ~ U (a, b)

**Example 1**

The average weight gained by a person over the winter months is uniformly distributed and ranges from 0 to 30 lbs. Find the probability of a person that he will gain between 10 and 15lbs in the winter months.

**Solution:**

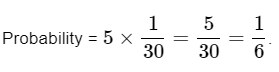
First, find the total height of the distribution.

The area under the probability distribution is always 1. Since there are 30 units starting from 0 to 30) the height is 

Then find the width of the slice of the distribution. Do this with subtracting the biggest number b from the smallest number a and you will get

**b – a = 15 – 10 = 5.**

Then multiply the width in Step 2 by the height in Step 1 and you will get

****

**Example 2:** Determine P(X ≤ 10) for the above-given question.

**Solution:**

This question is asking you to find the probability which the random variable X is lesser than 10. In simpler words, you need to determine the probability of the person gaining up to ten pounds.

Find the width of the box first which is b – a = 10 – 0 = 10.

Then multiply the width in Step 1 by the height. You already know that the height is 



**The Uniform distribution function used in R**

|  |  |
| --- | --- |
| **Function** | **Description** |
| dunif() | Returns the density |
| punif() | Calculate the uniform cumulative distribution function. This is, the probability of a variable X taking a value lower than x (that is, x <= X). If we need to compute a value x > X, we can calculate 1 – punif(x). |
| qunif() | Returns the quantile function |
| runif() | generate a sequence of random |

**Ex1.** Generate 15 random numbers between 1 and 3

**Solution: R-code with output**

|  |
| --- |
| > print("Random 15 numbers between 1 and 3")  **[1] "Random 15 numbers between 1 and 3"**  > runif(15, min=1, max=3)  **[1] 2.347028 2.492238 1.083744 2.510764 1.999683 2.360536 1.346497 1.049376**  **[9] 1.436326 1.825402 1.667491 2.943702 1.255876 1.237192 2.850865**  > rand\_unif<-runif(15, min=1, max=3)  > hist(rand\_unif, freq = FALSE, xlab = "x", density = 20, col = "darkgray") |

**Ex2.** Further, we plot both, the density histogram from above, as well as the uniform probability distribution for the interval [-2, 0.8], by applying the dunif() function:

**Solution: R-code with output**

|  |
| --- |
| > rand\_unif<-runif(10000, min=-2, max=0.8)  > hist(rand\_unif,freq = FALSE,xlab = "x",ylim = c(0, 0.4),xlim = c(-3, 3),  + density = 20,main = "Uniform distribution for the interval [-2,0.8]",col = "darkgray")  > curve(dunif(x, min = a, max = b),  + from = -5, to = 5,  + n = 100000,  + col = "darkblue",  + lwd = 2,  + add = TRUE,  + yaxt = "n",  + ylab = "probability"  + ) |

**Ex3. Calculating Punif value**

|  |
| --- |
| > # calculating punif value  > punif (15 , min =min , max = max)  **[1] 0.25** |

* 1. **Student’s t-distribution**

As we know normal distribution assumes two important characteristics about the dataset: a large sample size and knowledge of the population standard deviation. However, if we do not meet these two criteria, and we have a small sample size or an unknown population standard deviation, then we use the t-distribution.

**What is t-Distribution?**

Student’s t-distribution, also known as the t-distribution, is a probability distribution that is used in statistics for making inferences about the population mean when the sample size is small or when the population standard deviation is unknown. It is similar to the standard normal distribution (Z-distribution), but it has heavier tails. Theoretical work on t-distribution was done by W.S. Gosset; he has published his findings under the pen name “Student“. That’s why it is called a Student’s t-test. The t-score represents the number of standard deviations the sample mean is away from the population mean.

**The Formula for t-Distribution**

**t = [x̄-μ] / [s/sqrt(n) ]**

where,

**t** = the t-score,

**x̄** = sample mean,

**μ** = population mean,

**s** = standard deviation of the sample,

**n** = sample size

**When to Use the t-Distribution?**

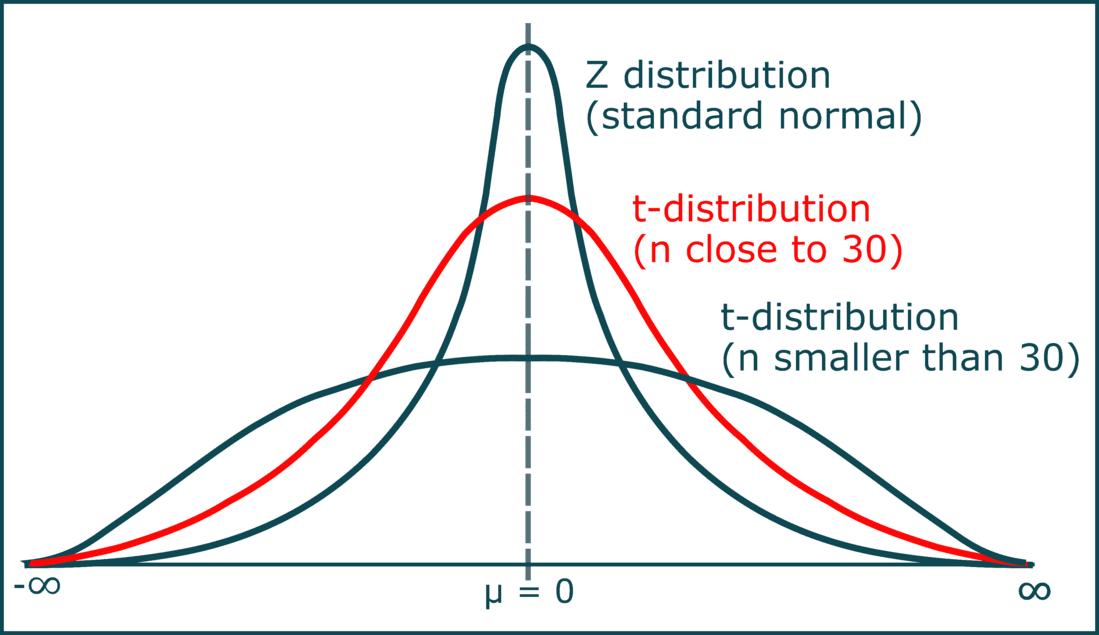
Student’s t Distribution is used when

* The sample size is 30 or less than 30.
* The population standard deviation (σ) is unknown.
* The population distribution must be unimodal and skewed.

**Properties of the t-Distribution**

The variable in t-distribution ranges from -∞ to +∞ (-∞ < t < +∞).

* t- distribution will be symmetric like the normal distribution if the power of t is even in the probability density function (pdf).
* For large values of ν(i.e. increased sample size n); the t-distribution tends to a standard normal distribution. This implies that for different ν values, the shape of t-distribution also differs.
* The t-distribution is less peaked than the normal distribution at the center and higher peaked in the tails. From the above diagram, one can observe that the red and green curves are less peaked at the center but higher peaked at the tails than the blue curve.
* The value of y(peak height) attains highest at μ = 0 as one can observe the same in the above diagram.
* The mean of the distribution is equal to 0 for ν > 1 where ν = degrees of freedom, otherwise undefined.
* The median and mode of the distribution is equal to 0.
* The variance is equal to ν / ν-2 for ν > 2 and ∞ for 2 < ν ≤ 4 otherwise undefined.



**Example:**

**Question**

There is a class of 25 students and the mean score of their test is 60 out of 100, with standard deviation 4 marks from the mean. While other students of the school have a mean score of 50 on the same test. What will be the t-score for calculating the probability that school students scored not less than 60 in their tests?

**Solution**

Let us begin assembling the values given in the question. From the question we can infer that the sample here is the class students and the population consists of all the students in the school.

The samples size of the class (N) - 25

Mean score of the class (x̅) - 60

Mean score of the population ( μ) - 50The standard deviation of the sample (s) - 4

Since we have got all the values that are required to calculate the t-score, we can simply insert them in the formula below

t = ( x̅ - μ) ÷ (s / √N),

t= (60 - 50) ÷ (4 / √25)

t= 10 ÷ 0.8

t= 12.5

The t-value obtained here leads to the cumulative probability from the t-distribution table from where you can find the log value of this t-score with the degrees of freedom, the sample means, the population means and standard deviation for this sample.

**Note:** *Student’s t table is a reference table that lists critical values of t. Student’s t table is also known as the t table, t-distribution table, t-score table, t-value table, or t-test table.*

**Student’s t- distribution functions**

|  |  |  |
| --- | --- | --- |
|  | **Purpose** | **Syntax** |
| **rt** | Generates random numbers from t-distribution. Given number of observations (n) and the degrees of freedom (df). | rt(n, df) |
| **dt** | Density of t-distribution. Given vector of quantiles (x) and the degrees of freedom (df). | dt(x, df) |
| **pt** | Cumulative Distribution Function (CDF). Given a vector of probabilities (p), the degrees of freedom (df) and (lower.tail) which is set to true for a less than calculation or false for a greater than calculation. | pt(q, df, lower.tail) |
| **qt** | Quantile Function. Given vector of quantiles (q) and the degrees of freedom (df) | qt(p, df) |

**Note:** *The degrees of freedom represent the number of values in a statistical calculation that are free to vary. In the case of the t-distribution, the degrees of freedom are N-1 as one degree of freedom is reserved for estimating the mean, and N-1 degrees remain for estimating the variability.*

1. **dt()**

|  |
| --- |
| > # Set degrees of freedom  > df <- 5  > # Create a sequence of x values  > x <- seq(-5, 5, 0.01)  > # Calculate the density  > density <- dt(x, df)  > # Plot the density  > plot(x, density, type="l", main="t-Distribution Density", xlab="x", ylab="Density") |

1. **pt()**

|  |
| --- |
| *# Set degrees of freedom*  df <- 5  *# Create a sequence of x values*  x <- seq(-5, 5, 0.01)  *# Calculate the cumulative distribution function*  cdf <- pt(x, df)  *# Plot the cumulative distribution function*  plot(x, cdf, type="l", main="t-Distribution CDF", xlab="x", ylab="CDF") |

1. **qt()**

|  |
| --- |
| > # Set degrees of freedom  > df <- 5  > # Create a sequence of p values  > p <- seq(0, 1, 0.01)  > # Calculate the quantile function  > quantiles <- qt(p, df)  > # Plot the quantile function  > plot(p, quantiles, type="l", main="t-Distribution Quantile Function", xlab="p", ylab="Quantile") |

1. **rt()**

|  |
| --- |
| > # Set degrees of freedom  > df <- 5  > n <- 1000  > # Generate random variates  > random\_variates <- rt(n, df)  > # Plot the histogram of the random variates  > hist(random\_variates, main="Histogram of Random Variates from t-Distribution", xlab="Value", ylab="Frequency") |

**Ex.** Plotting the graph with multiple degrees of freedom

|  |
| --- |
| > abline(v = 0, lty = 2, col = "black")  > # Define a range of x values  > x <- seq(-5, 5, by = 0.01)  > # Define a vector of degrees of freedom  > df\_values <- c(1, 5, 20)  > # Plot the t-distribution curves for different degrees of freedom  > colors <- c("red", "blue", "green")  > legend\_labels <- c("df = 1", "df = 5", "df = 20")  > # Create an empty plot  > plot(x, dt(x, df = df\_values[1]), type = "l", col = colors[1], ylim = c(0, 0.4), xlab = "x", ylab = "Density", main = "t-Distribution with Different Degrees of Freedom")  > # Add curves for different degrees of freedom  > for (i in 2:length(df\_values)) { lines(x, dt(x, df = df\_values[i]), col = colors[i]) }  > # Add a legend  > legend("topright", legend = legend\_labels, col = colors, lty = 1)  > # Add a vertical line at x = 0  > abline(v = 0, lty = 2, col = "black") |