Probability Distribution

Probability Distribution is a function that shows all the possible values a variable can take and how often it occurs.

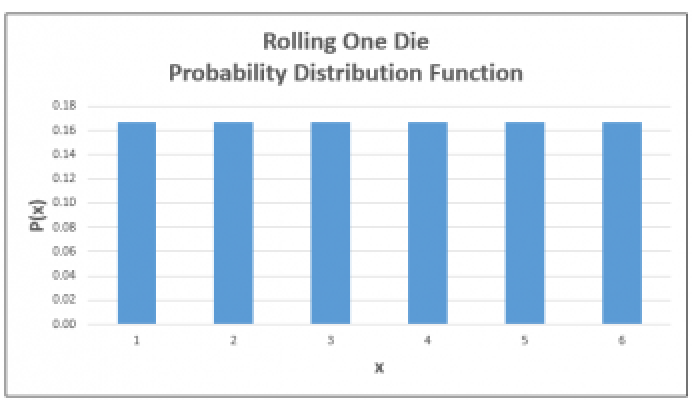
Probability Distribution is defined by the underlying probability and graph is just the visual representation.

Let's take an example and find out probability distribution of all possible outcomes of rolling a die:

Possible Outcomes	Probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

The above table represents the probability distribution of the variable (possible outcomes).

When we plot the above table, we get the probability distribution graph as below.



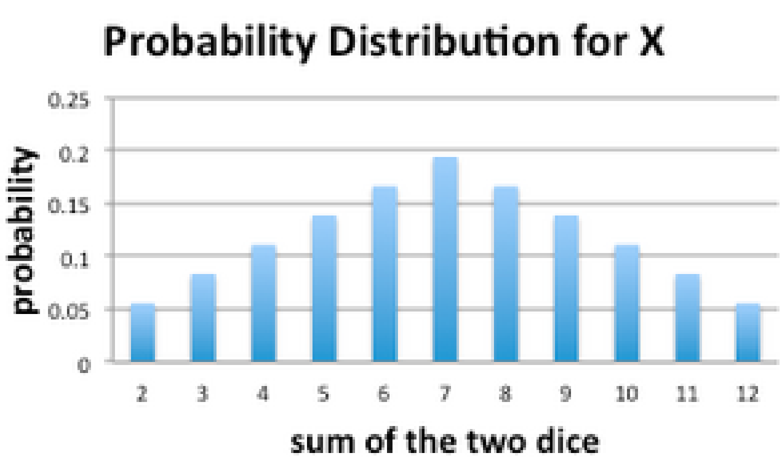
Let's take another example and find the probability distribution of getting sums of rolling two dice: -

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

Number of outcomes = 36

Possible Sums	Occurrence	Probability
2	1	0.03
3	2	0.06
4	3	0.08
5	4	0.11
6	5	0.14
7	6	0.17
8	5	0.14
9	4	0.11
10	3	0.08
11	2	0.06
12	1	0.03

This table represents the Probability distribution of different outcomes and if we plot the above table, we get the below Probability distribution graph.



Probability Distributions: Discrete vs. Continuous

Variables can be of two types discreet and continuous.

For example, when we toss a coin or roll a die, we know the possible outcome will be 1 head, 2 tails or in case of die 1,2 ,6 etc. We can never get values like 1.5,3.2. Such variables are types of **Discreet variables**.

On the other hand, let's say criteria for selection in basket ball team is that one should have a height between 170 cm to 200 cm. Here height will be a type of **continuous variable** as one can take any value between 170 and 200.

Discrete Probability Distribution

If a random variable is discreet, its probability distribution is said to be Discrete Probability Distribution.

Different types of Continuous Probability Distribution

1) Binomial Distribution
2) Poisson's distribution
3) Bernoulli's distribution

The examples we took above for rolling a die is a type of Discrete Probability Distribution.

Continuous Probability Distribution

Similarly, when a random variable is continuous, its probability distribution is said to be continuous Probability Distribution.

In a continuous probability distribution, unlike a discrete probability distribution, the probability that a continuous random variable will assume a particular value is zero. Thus, we cannot express a continuous probability distribution in tabular form. We describe it using an equation or a formula also known as Probability Density Function (pdf).

For a continuous probability distribution, the probability density function has the following properties:

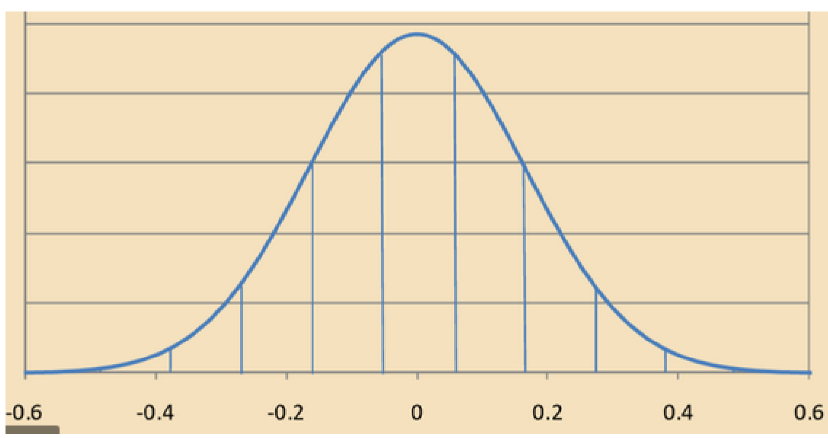
- The graph of the density function will always be continuous over the range in which the random variable is defined.
- The area bounded by the curve of the density function and the x-axis is equal to 1, when computed over the domain of the variable.
- The probability that a random variable assumes a value between a and b is equal to the area under the density function bounded by a and b.

Different types of Continuous Probability Distribution

4) Normal Distribution
5) Student's T distribution
6) Chi Squared distribution

Normal Distribution

Normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve (as in the figure below).



The probability density function for Normal distribution is given as:

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

where

- μ is the mean or expectation of the distribution (and also its median and mode)
- σ is the standard deviation
- σ^2 is the variance

Given a variable x(a), the probability of the random variable X, which follows a normal distribution, is less than or equal to x(a) is given as:

$$Pr\{X \leq x_a\} = \int_{-\infty}^{x_a} f(x)dx = \int_{-\infty}^{x_a} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx,$$

Here we integrate the **pdf** function to get the cumulative probability.

<u>**Note**</u>: For all the below examples we will use the "Normal Distribution Calculator" as it is very easy for calculation purposes. We can do the calculations with the above given formula also but it's not necessary. We will learn how to solve problems using z-score table in the next section.

Examples:

Q) The Light Bulb Company has found that an average light bulb lasts 900 hours with a standard deviation of 80 hours. Assuming that bulb life is normally distributed. What is the probability that a randomly selected light bulb will burn out in 1000 hours or less?

Answer:

Mean = 900

Standard deviation = 80

x(a) = 1000

we have to use the above given formula for finding the probability. But for the time being we are taking help of the Normal distribution calculator (shown below).

Normal random variable (x)	1000
Cumulative probability: P(X ≤ 1000)	0.894
Mean	900
Standard deviation	80

Putting the above values, we get a probability of:

P(X<=1000) = 89.4% i.e. there is 89.4% chance that the bulb will burnout within 1000 hours.

Q) Suppose scores on a mathematics test are normally distributed. If the test has a mean of 55 and a standard deviation of 10, what is the probability that a person who takes the test will score between 45 and 65?

Ans.

Here we need to find the probability for 45<= x(a) <=65.

If we find the cumulative probability for x<=45 and cumulative probability for x<=65, we can subtract them to find the required probability.

Mean = 55

Standard deviation = 10

Again, using the normal distribution calculator, we find both the values as:

P(x(a)< = 45) = 0.159

P(x(a)< = 65) = 0.841

So, P (45< X(a) <65) = 0.841 – 0.159

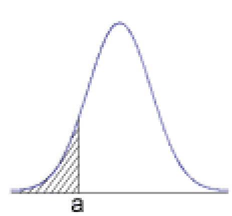
= 0.682

So, we have 68.2% probability of a person taking test to score between 45 and 65 marks

----

The Normal Distribution has following properties:

1) mean = median = mode
2) symmetry about the center
3) 50% of values less than the mean and 50% greater than the mean
4) The probability that X is greater than ‘a’ is equal to the area under the normal curve as shown by the non-shaded area in the figure below.
5) The probability that X is less than ‘a’ is equal to the area under the normal curve as shown by the shaded area in the figure below.

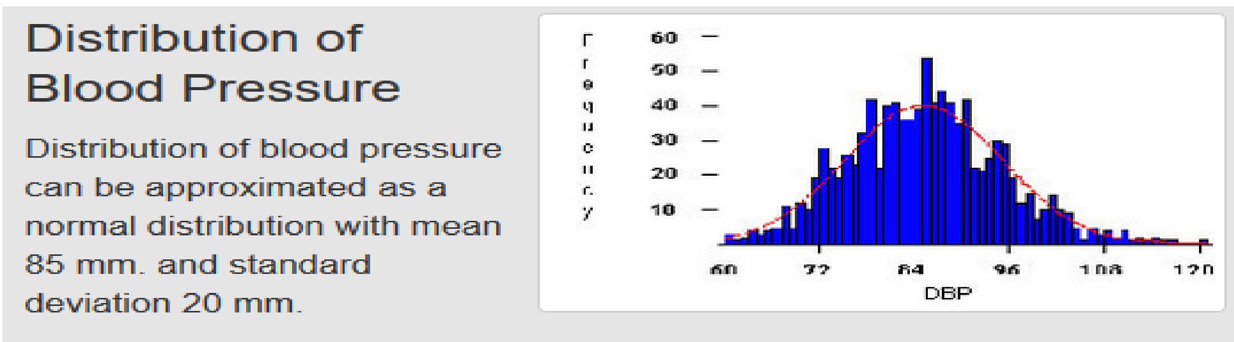


Let's see some real-life examples which follows normal distribution

When we talk about the heights or weights of people in the world, it is seen that it follows a normal distribution. Why does it seem obvious? It is simple because it is more probable to find people with heights near the average rather finding very short heighted or very tall people. Just look around in your class, you will find the majority of people will fall in the range near the average height of the class.

Similar is the case with the weights and IQ of people.

Distribution of wealth is another example of Normal Distribution. Most of the people fall in the average wealth category. ("Middle class").



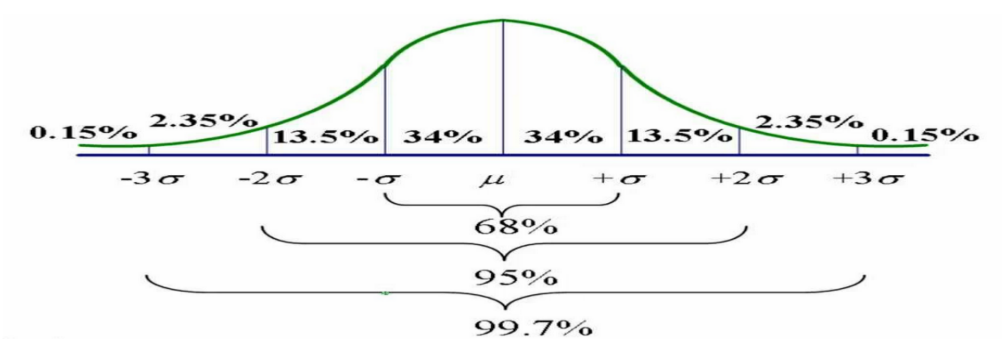
<u>**The empirical rule (Three Sigma Rule)**</u>

It states that for a Normal Distribution, nearly all of the data will fall within three range of standard deviations of the mean. The empirical rule can be understood when broken down into three parts:

1) 68% of the data falls within the first standard deviation from the mean.

2) 95% fall within two standard deviations.

3) 99.7% fall within three standard deviations.



Why do we use Normal Distribution?

1) Distributions of sample means with large sample sizes can be approximated t0 normal distribution
2) Decisions based on normal distribution insights have proven to be of good value
3) All computable statistics are elegant
4) It approximates a wide variety of random variables

Standard Normal Distribution

The standard normal distribution is a special case of the normal distribution. It is the distribution that occurs when a normal random variable has a **mean of zero** and a **standard deviation of one**.

The normal random variable of a standard normal distribution is called a **standard score** or a **z score**.

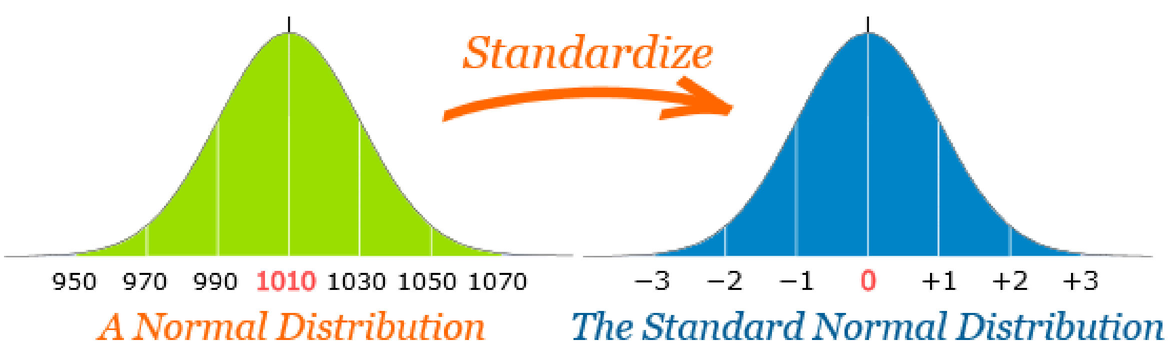
Every normal random variable X can be transformed into a z score via the following equation:

$z = (X - \mu) / \sigma$

where, X is a normal random variable,

μ is the mean,

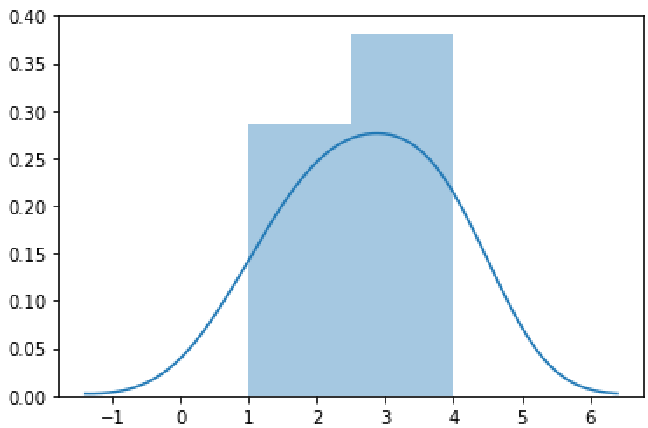
and σ is the standard deviation.



Steps to transform into Standard N-Distribution

Suppose we have a dataset with elements

X = [1,2,2,3,3,4,4]



we can see the data is normally distributed.

steps:

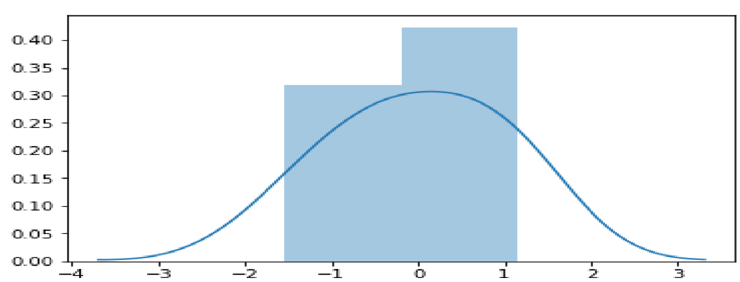
1) mean= 2.71

2) std_deviation = 1.11

3) transform to z scores using z= (x-mean)/std_deviation

4)After transforming X = {-1.54,-0.64,-0.64,0.26,0.26,1.16,1.16}

Plotting new data:



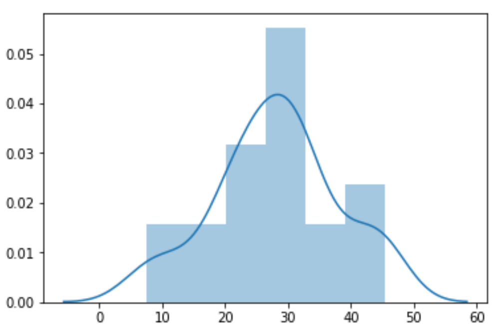
mean=0, std_deviation=1

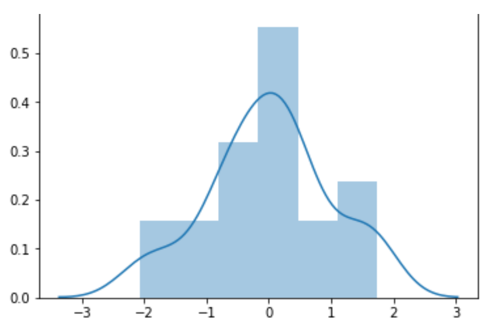
<u>Let's take another example and see how the transformation works:</u>

Given a data set , we were provided with the initial_data , we transformed the data and found the z score as shown in the below table.

Initial_ Data	Initial_data - mean	z-score =(Initial_data - mean)/St. Deviation
45.363	17.295	1.727449369
33.435	5.367	0.536092603
17.713	-10.355	-1.034292115
26.019	-2.049	-0.204671347
27.508	-0.560	-0.055885355
11.194	-16.874	-1.685409951
19.894	-8.174	-0.816445735
23.133	-4.935	-0.492868567
29.126	1.058	0.105720916
30.427	2.359	0.235629855
44.729	16.661	1.664114735
7.393	-20.675	-2.065026984
26.983	-1.085	-0.108386711
30.616	2.548	0.254473261
32.662	4.594	0.458842761
23.013	-5.055	-0.504921837
30.949	2.881	0.287795782
37.189	9.121	0.910991079
21.990	-6.078	-0.607100949
42.025	13.957	1.393988737
mean =28.068		mean= 0
St. deviation =10.012		St. Deviation= 1

Let's see the plots for both the initial data and transformed data:





Initial data

Transformed data

We can see from the graphs that our normal distribution when transformed has zero mean and deviation of 1.

z score

z-score is a measure of position that indicates the number of standard deviations a data value lies from the mean. z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

z score is very powerful tool to find the probability distribution using the z-score table.

We do not need the normal distribution calculator and will simply use z-score table.

x	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

z-score table

let’s see the application of z-score table and see how to use it:

We will use the same example from above where we used the Normal Distribution calculator, let’s try to calculate the same using z-score table and compare the results.

Q) The Light Bulb Company has found that an average light bulb lasts 900 hours with a standard deviation of 80 hours. Assuming that bulb life is normally distributed. What is the probability that a randomly selected light bulb will burn out in 1000 hours or less?

Answer: let's convert our data in standard normal form.

Mean = 900

Std. deviation = 80

x(a) = 1000

standardized x(a) = (1000-900)/80 = 1.25

z-score = 1.25 = 1.2 + 0.05 (in the table we will match the value corresponding to 1.2 and 0.05)

Let's use the z-score table for this:

X	0.0	0.01	0.02	0.03	0.04	0.05	
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	(
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	(
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	(
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	(
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	(
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	(
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	(
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	(
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	(
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	(
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	(
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	(
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	(
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	(

We found that that Probability coming as 89.44 %.

This is exactly same as we found out with the Normal distribution calculator.

Thus, if we standardize a normal distribution, z-score becomes a very important tool in helping finding the probability distribution.

Q) Ravi scored 980 in a Physics Olympiad. The mean test score was 870 with a standard deviation of 120. How many students scored more than Ravi? (Assume that test scores are normally distributed.)

Answer: Let's standardize the test score

Mean = 870

St. deviation= 120

z-score = (980-870)/120 = 0.917 (we will approximate it to 0.92)

X	0.0	0.01	0.02	0.03
0.0	0.5000	0.5040	0.5080	0.5120
0.1	0.5398	0.5438	0.5478	0.5517
0.2	0.5793	0.5832	0.5871	0.5910
0.3	0.6179	0.6217	0.6255	0.6293
0.4	0.6554	0.6591	0.6628	0.6664
0.5	0.6915	0.6950	0.6985	0.7019
0.6	0.7257	0.7291	0.7324	0.7357
0.7	0.7580	0.7611	0.7642	0.7673
0.8	0.7881	0.7910	0.7939	0.7967
0.9	0.8159	0.8186	0.8212	0.8238

So, P(x<=980) = 0.8212

We need to find the probability of scoring more than 980,

P(x>980) = 1 – 0.8212 = 0.1788

Thus, we can estimate 17.88% students scored more than Ravi in the test.

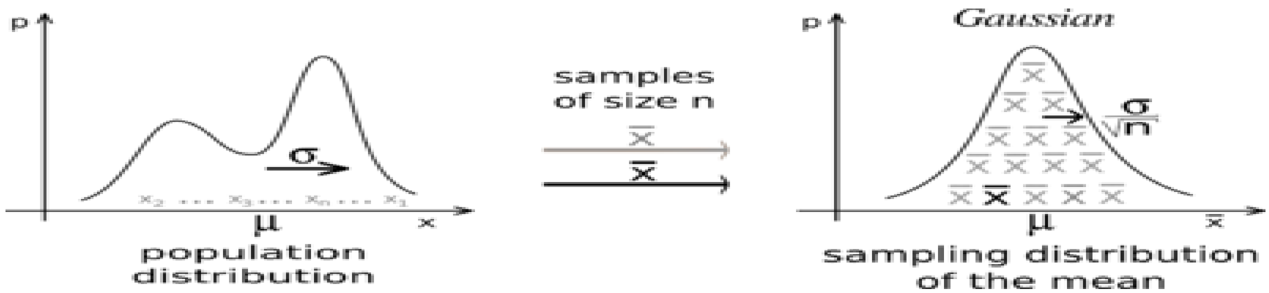
Note: When you encounter a negative z score, you can use a negative z score table, or find the value for positive value of the z-score and subtract it from 1.

e.g. p (-2.5) = 1 – p (2.5) = 1 – 0.99379 = 0.00621, which same as the value you will find in negative z score table.

Central Limit Theorem

In the study of probability theory, the central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution (also known as a “bell curve”), as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population distribution shape.

CLT is a statistical theory stating that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population. Furthermore, all the samples will follow an approximate normal distribution pattern, with all variances being approximately equal to the variance of the population, divided by each sample's size. The samples extracted should be bigger than 30 observations.



Let’s visualize the CLT with few data sets with different sample sizes:

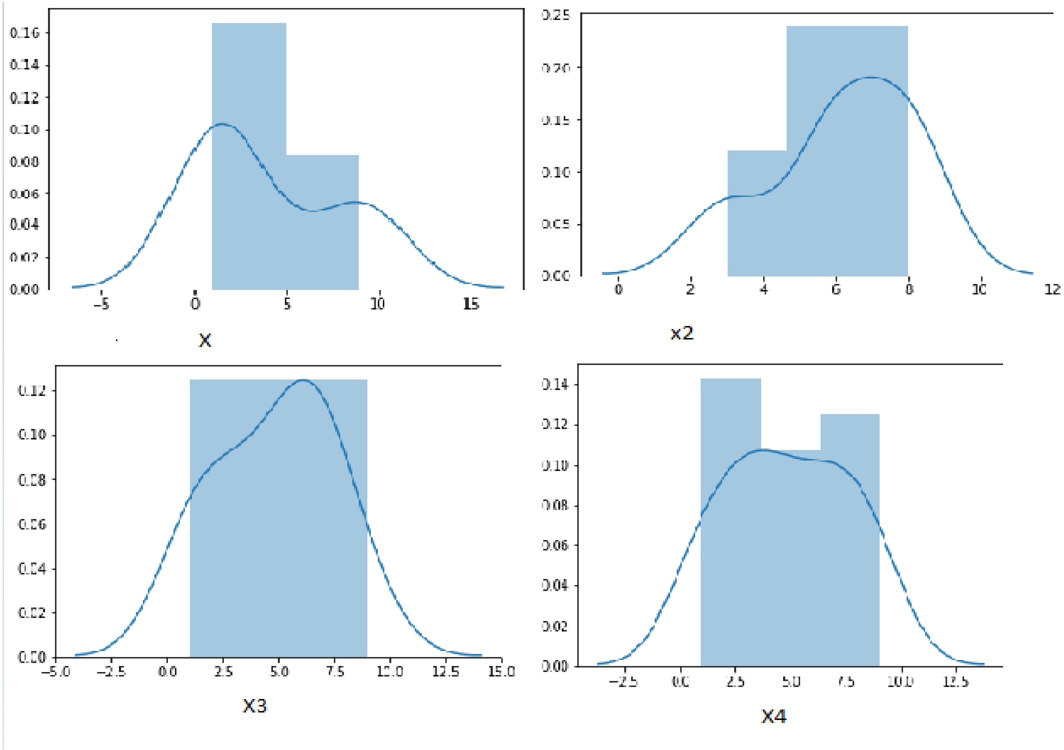
x= [9,2,1]

x2= [6,6,8,3,8]

x3= [5,3,6,4,7,2,6,9,7,1,1,7]

x4 = [8,1,7,1,4,3,1,7,8,9,8,3,1,6,8,3,4]

Plotting the distribution graph for above samples with different sample size



We can clearly see that among the number of samples extracted from the population, as the sample size increases, sample moves closer to a Normal Distribution.

We will talk more about Central limit theorem and see it's uses, but first let's discuss one more important concept.

Standard Error

The standard error (SE) of a statistic is the approximate standard deviation of a statistical sample population. The standard error is a statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation. In statistics, a sample mean deviates from the actual mean of a population—this deviation is the standard error of the mean.

When a population is sampled, the mean, or average, is generally calculated. **The standard error can include the variation between the calculated mean of the population and one which is considered known, or accepted as accurate. This helps compensate for any incidental inaccuracies related to the gathering of the sample.**

In cases where multiple samples are collected, the mean of each sample may vary slightly from the others, creating a spread among the variables. This spread is most often measured as the standard error, accounting for the differences between the means across the datasets.

The more data points involved in the calculations of the mean, the smaller the standard error tends to be. When the standard error is small, the data is said to be more representative of the true mean. In cases where the standard error is large, the data may have some notable irregularities.

The standard deviation is a representation of the spread of each of the data points. The standard deviation is used to help determine the validity of the data based on the number of data points displayed at each level of standard deviation. Standard errors function more as a way to determine the accuracy of the sample or the accuracy of multiple samples by analyzing deviation within the means.

Standard Error is given by the following formula:

Standard Error

$$Standard\ Error\ (\sigma_{\mu}) = \frac{\sigma}{\sqrt{n}}$$

$$\sigma = standard\ deviation$$
$$n = quantity\ of\ numbers\ in\ the\ group$$

Here σ is the standard deviation of the population, whereas σ(u) is the standard deviation of the sample.

We can see that as the size of our sample increases the Standard error decrease.

Now, that we know the Standard error, let's rephrase our Central Limit theorem as:

The central limit theorem states that the sample mean follows approximately the normal distribution with mean(μ) and standard deviation (σ/√n), where μ and σ are the mean and standard deviation of the population from where the sample was selected. The sample size n has to be large (usually n≥30) if the population from where the sample is taken is non normal.

So, when we transform our sample data, we will use following formula for the z-score:

z = (X - μ) / (σ/√n)

where, X is the sample mean,

μ is the mean of the population,

and σ is the standard deviation of the population.

Let's see an example based on the above explaination.

Q) Let X be a random variable with μ= 10 and σ= 4. A sample of size 100 is taken from this population. Find the probability that the sample mean of these 100 observations is less than 9.

Ans: population mean = 10 population std. deviation = 4 sample size(n) = 100

Sample mean = 9

z= (9-10)/ (4/ (100) ^0.5) = -2.5

We will use the z-score table and find the value to be 0.0062

P(X<9) = 0.0062

Q) A large freight elevator can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean= 205 pounds and standard deviation = 15 pounds. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?

Ans: For all the boxes to be loaded the total weight must be at most 9800.

So, the sample mean should be = 9800/49 = 200 sample size(n) = 49

Population mean = 205

Std. deviation = 15

z-score = (200-205)/(15/(49)^0.5) = -2.33

using z-score table:

z	.00	.01	.02	.03	.04	.05	.06	.07
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307

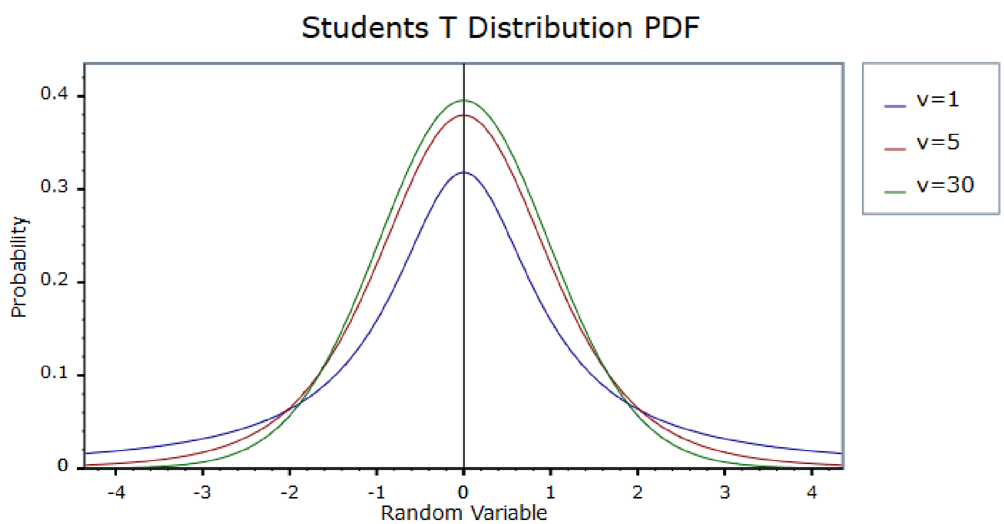
P(X<200) = 0.0099

Student's t-Distribution

The t distribution (aka, Student's t-distribution) is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.

The t distribution is very similar to the normal distribution when the estimate of variance is based on many degrees of freedom, but has relatively more scores in its tails when there are fewer degrees of freedom.

The t distribution approaches the normal distribution as the degrees of freedom increase.



According to the central limit theorem, the sampling distribution of a statistic (like a sample mean) will follow a normal distribution, as long as the sample size is sufficiently large. Therefore, when we know the standard deviation of the population, we can compute a z-score, and use the normal distribution to evaluate probabilities with the sample mean.

But sample sizes are sometimes small, and often we do not know the standard deviation of the population. When either of these problems occur, statisticians rely on the distribution of the t statistic (also known as the t score), whose values are given by:

$t = (x- \mu) / (s/\sqrt{n})$

where x is the sample mean,

μ is the population mean,

s is the standard deviation of the sample,

and n is the sample size.

The distribution of the t statistic is called the t distribution or the Student t distribution.

The t distribution allows us to conduct statistical analyses on certain data sets that are not appropriate for analysis, using the normal distribution.

Degrees of Freedom

There are actually many different t distributions. The particular form of the t-distribution is determined by its degrees of freedom. The degrees of freedom refer to the number of independent observations in a set of data.

When estimating a mean score or a proportion from a single sample, the number of independent observations is equal to the sample size minus one. Hence, the distribution of the t statistic from samples of size 8 would be described by a t distribution having 8 - 1 or 7 degrees of freedom. Similarly, a t distribution having 15 degrees of freedom would be used with a sample of size 16.

T score table

Table C: T-distribution Chart

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

EXAMPLE:

Q) A lock Company claims that their locks have an average breaking strength of 1000 Kg, with a standard deviation of 150 kg. Suppose a customer tests 8 randomly-selected locks. What is the probability that the average breaking strength in the test will be no more than 800 Kg?

Answer: population mean= 1000

sample mean= 800

std. deviation= 85

n= 8

t-score = (800 - 1000) / (150/ (8) ^0.5) = -3.77

We will use t-distribution calculator for this exercise:

Random variable	t score
Degrees of freedom	7
t score	-3.77
Probability: $P(T \leq -3.77)$	0.0035

We get a cumulative probability for P(X<800) = 0.0035

Note: This was a very simple case of t-distribution. In real life, t-test is one of the most important and most used tests in Hypothesis testing. There are many terms associated such as critical value, significance, 1 tail test and 2 tail test. We will study more in detail about t-tests and use of t-score in upcoming topic "Hypothesis Testing".

Bernoulli Distribution

It is a type of Discrete Probability distribution. The Bernoulli distribution essentially models a single trial of flipping a weighted coin. It is the probability distribution of a random variable taking on only two values, 1 ("success") and 0 ("failure") with complementary probabilities p and 1−p respectively. The Bernoulli distribution therefore describes events having exactly two outcomes, which are present in real life.

Suppose We have a single trial of with only two possible outcomes success or failure:

P(Success) = p

P(Failure)= 1-p

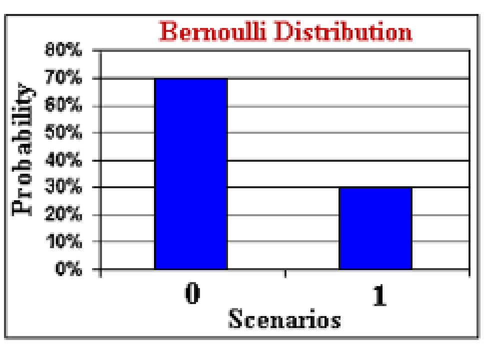
Let, X=1 when Success and X=0 when failure,

Then the probability distribution function is given as:

$P(x) = p^x(1-p)^{1-x}.$ for x= (0,1)

So, P(1) = p^1 * (1-p) ^0 = p   P(0) = P^0*(1-p) ^1 = 1-p

A simple graphical representation of Bernoulli's distribution will look like this.



Here, p =0.3

The Expected value (mean) for the Bernoulli's distribution is given as:

$$E(X) = 0 \times (1-p) + 1 \times p = p.$$

The variance for the Bernoulli's distribution is given as:

$$Var(X) = E(X^2) - E(X)^2 = 1^2 \times p + 0^2 \times (1-p) - p^2 = p - p^2 = p(1-p).$$

Some real-life cases that follows a Bernoulli distribution:

1) Results of Exam (Pass or Fail)
2) Gender of New born baby (Male or Female)
3) Result of Cricket World Cup (Win or Lose)
4) Tossing a coin (Heads or Tails)

We will see some more examples when Bernoulli trial is repeated for many times.

Binomial Distribution

A binomial experiment is a series of n Bernoulli trials, whose outcomes are independent of each other. A random variable, X, is defined as the number of successes in a binomial experiment.

For example, consider a fair coin. Flipping the coin once is a Bernoulli trial, since there are exactly two complementary outcomes (flipping a head and flipping a tail), and they are both 1/2 no matter how many times the coin is flipped. Note that the fact that the coin is fair is not necessary; flipping a weighted coin is still a Bernoulli trial.

A binomial experiment might consist of flipping the coin 100 times, with the resulting number of heads being represented by the random variable X. The binomial distribution of this experiment is the probability distribution of X.

If **X** is the number of success in a given Bernoulli trial with **n** independent trials, with probability of success being **p** and probability of failure being **1-p**, then for exactly k success in the experiment, the probability distribution is given as:

P(X=k) = $\binom{n}{k} p^k (1-p)^{n-k}.$

Here (n k) is the numner of ways of choosing k from n also written as C(n,k).

Q) Let's flip a coin 6 times with probability of getting a tail be 0.3. Let's write the binomial distribution for this experiment.

Ans:

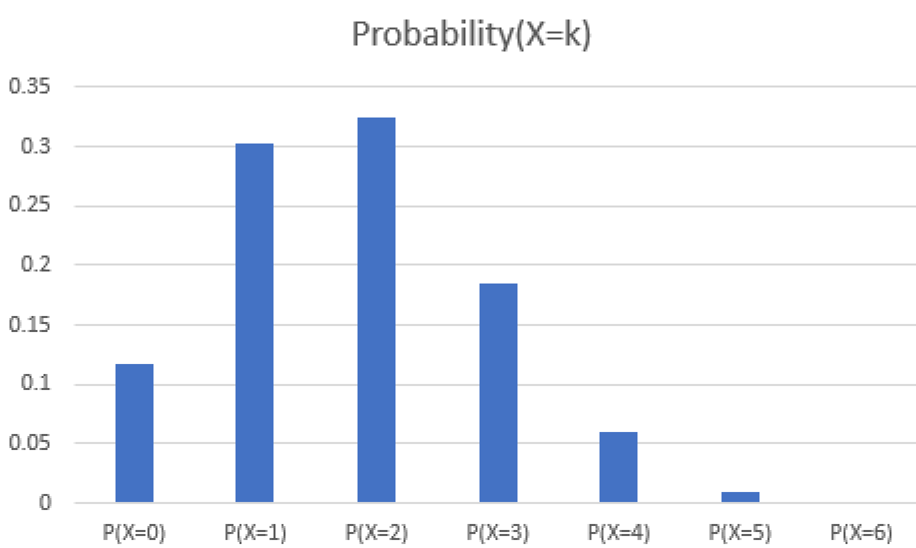
Possible outcomes(k)	Probability(X=k)	Binomial Distribution	Probability(X=k)
0 tail	P(X=0)	C(6,0) * (0.30^0)*(0.70)^6	0.118
1 tail	P(X=1)	C(6,1) * (0.30^1)*(0.70)^5	0.302
2 tail	P(X=2)	C(6,2) * (0.30^2)*(0.70)^4	0.324
3 tail	P(X=3)	C(6,3) * (0.30^3)*(0.70)^3	0.185
4 tail	P(X=4)	C(6,4) * (0.30^4)*(0.70)^2	0.06
5 tail	P(X=5)	C(6,5) * (0.30^5)*(0.70)^1	0.01
6 tail	P(X=6)	C(6,6) * (0.30^6)*(0.70)^0	0.0007
**How to calculate C(6,1) = 6!/[(6-1)! * 1!]			

So looking at the above table we can find probability of obtaining k outcomes.

e.g. what is the probability of getting exactly 2 tails in above experiment

P(X=2) = 0.324

Let's plot the above table,

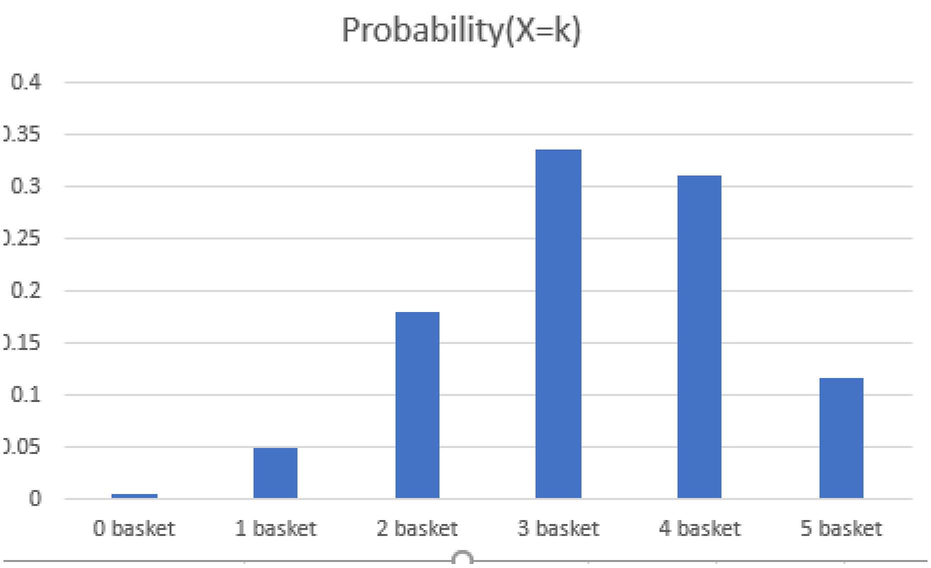


Q) A basketball player takes 5 independent free throws with a probability of 0.65 of getting a basket on each shot. Let X=the number of baskets he gets. Show the probability distribution for X.

Solution:

Possible outcomes(k)	Probability(X=k)	Binomial Distribution	Probability(X=k)
0 basket	P(X=0)	C(5,0) * (0.65^0)*(0.35)^5	0.0052
1 basket	P(X=1)	C(5,1) * (0.65^1)*(0.35)^4	0.049
2 baskets	P(X=2)	C(5,2) * (0.65^2)*(0.35)^3	0.181
3 baskets	P(X=3)	C(5,3) * (0.65^3)*(0.35)^2	0.336
4 baskets	P(X=4)	C(5,4) * (0.65^4)*(0.35)^1	0.312
5 baskets	P(X=5)	C(5,5) * (0.65^5)*(0.35)^0	0.116

Probability distribution Graph:



Mean and variance for Binomial Distribution

Mean= n*p

Variance = n* p(1-p),

where **n** is the number of trials, **p** is probability of success and **1-p** is probability of failure

Poisson Distribution

The Poisson distribution is the discrete probability distribution of the number of events occurring in a given time period, given the average number of times the event occurs over that time period.

Example

A certain car wash shop gets an average of 3 visitors to the center per hour. This is just an average, however. The actual amount can vary.

A Poisson distribution can be used to analyze the probability of various events regarding how many customers go to the center. It can allow one to calculate the probability of a dull activity (when there are 0 customers coming) as well as the probability of a high activity (when there are 5 or more customers coming). This information can, in turn, help the owner to plan for these events with staffing and scheduling.

If **X** is the number of events observed over a given time period, then probability of observing k events over the time period is:

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

The mean (or expected) rate is λ.

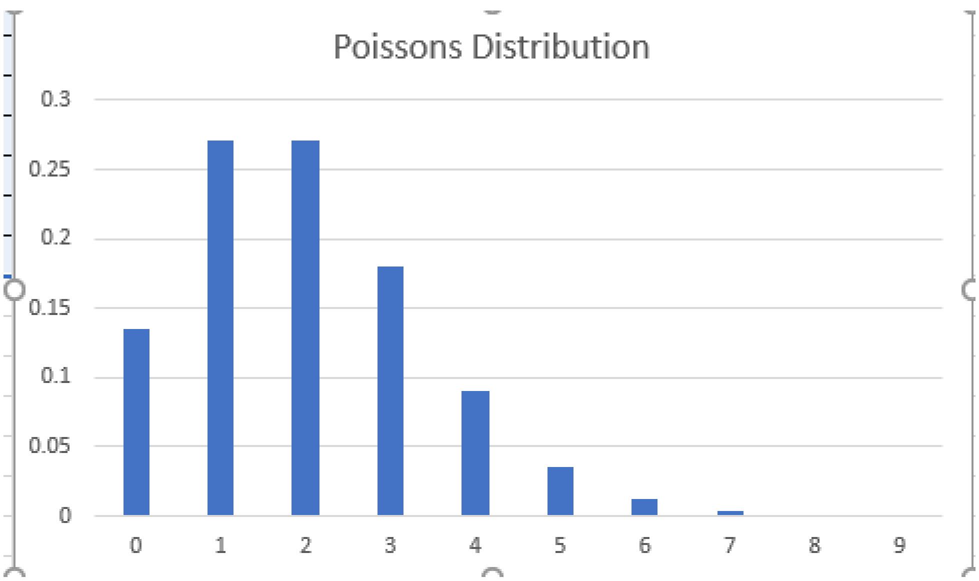
The Poisson distribution is often used as an approximation for binomial probabilities when n is large and p is small.

Q) In a coffee shop, average number of customers per hour is 2. Find the probability of getting k number of customers in the shop.

Number of customers	Probability(X=k)	Poissons Distribution	
0	P(X=0)	0.135335283	
1	P(X=1)	0.270670566	
2	P(X=2)	0.270670566	
3	P(X=3)	0.180447044	
4	P(X=4)	0.090223522	mean = 2
5	P(X=5)	0.036089409	
6	P(X=6)	0.012029803	
7	P(X=7)	0.003437087	
8	P(X=8)	0.000859272	
9	P(X=9)	0.000190949	

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

Let’s plot the probability distribution:



We can clearly see that probability of getting number customers starts declining after 6.

Q) Suppose the average number of elephants seen on a 1-day safari is 6. What is the probability that tourists will see fewer than 4 elephants on the next 1-day safari?

Solution:

Number of Elephants	Probability(X=k)	Poisson Distribution
0	P(X=0)	0.0025
1	P(X=1)	0.0149
2	P(X=2)	0.0446
3	P(X=3)	0.0892
4	P(X=4)	0.1339

Mean = 6

We need values $P(X<4) = P(X<=3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$

$= 0.0025+0.0149+0.0446+0.0892$

$= 0.1512$

Some applications that obey a Poisson distribution is below:

a. the number of mutations on a given strand of DNA per time unit
b. the number of bankruptcies that are filed in a month
c. the number of arrivals at a car wash in one hour
d. the number of network failures per day
e. the number of file server virus infection at a data center during a 24-hour period
f. the number of Airbus 330 aircraft engine shutdowns per 100,000 flight hours
g. the number of asthma patient arrivals in a given hour at a walk-in clinic
h. the number of hungry persons entering McDonald's restaurant per day
i. the number of work-related accidents over a given production time
j. the number of births, deaths, marriages, divorces, suicides, and homicides over a given period of time
k. the number of customers who call to complain about a service problem per month
l. the number of visitors to a web site per minute
m. the number of calls to consumer hot line in a 5-minute period
n. the number of telephone calls per minute in a small business