Mr.Gangadhar Immadi,

Immadi.gangadhar@gmail.com

9986789040

PROBABILITY DISTRIBUTIONS

Random Experiment

- Random experiment is an experiment in which the outcome is not known with certainty.

- Predictive analysis mainly deals with random experiment like:
  - Predicting quarterly revenue of an organization
  - Customer churn
  - Demand for a product at future time period etc.

Sample Space

- It is the universal set that consist of all possible outcomes of an experiment.

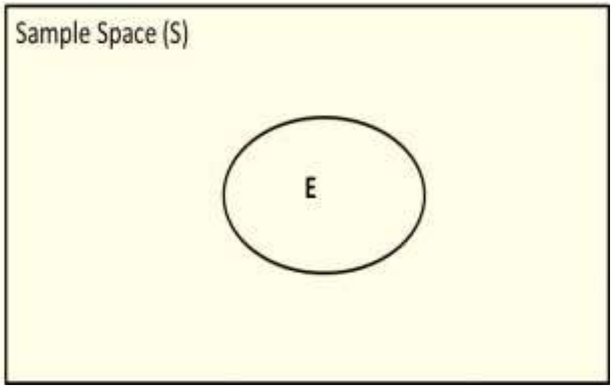
- It is represented using letter "S"

- Individual outcomes are called elementary events

- Sample Space can be finite or infinite.

Event

- Event(E) is a subset of a sample space and probability is usually calculated with respect to an event.



- The Venn diagram indicates that the event E is a subset of the sample space S, that is, $E \subset S$ (E is a subset of S)

Probability Estimation using Relative Frequency

- The classical approach to probability estimation of an event is based on the relative frequency of the occurrence of that event

- According to frequency estimation, the probability of an event X, P(X), is given by

$$P(X) = \frac{\text{Number of observations in favour of event } X}{\text{Total number of observations}} = \frac{n(X)}{N}$$

Example

A website displays 10 advertisements and the revenue generated by the website depends on the number of visitors to the site clicking on any of the advertisements displayed in the website. The data collected by the company has revealed that out of 2500 visitors, 30 people clicked on 1 advertisement, 15 clicked on 2 advertisements, and 5 clicked on 3 advertisements. Remaining did not click on any of the advertisements. Calculate

(a) The probability that a visitor to the website will click on an advertisement.
(b) The probability that the visitor will click on at least two advertisements.
(c) The probability that a visitor will not click on any advertisements.

Solution

(a) Number of customers clicking an advertisement is 50 and the total number of visitors is 2500. Thus, the probability that a visitor to the website will click on an advertisement is

$$\frac{50}{2500} = 0.02$$

(b) Number of customers clicking on at least 2 advertisements is 20. Thus, the probability that a visitor will click on at least 2 advertisements is

$$\frac{20}{2500} = 0.008$$

(c) Probability that a visitor will not click on any advertisement is

$$\frac{2450}{2500} = 0.98$$

Algebra of Events

- Assume that X, Y and Z are three events of a sample space. Then the following algebraic relationships are valid and are useful while deriving probabilities of events:

- **Commutative rule:** $X \cup Y = Y \cup X$ and $X \cap Y = Y \cap X$

- **Associative rule:** $(X \cup Y) \cup Z = X \cup (Y \cup Z)$ and $(X \cap Y) \cap Z = X \cap (Y \cap Z)$

- **Distributive rule:** $X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z)$

  $X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z)$

- *DeMorgan's Laws* on complementary sets are useful while deriving probabilities:

$$(X \cup Y)^C = X^C \cap Y^C$$

$$(X \cap Y)^C = X^C \cup Y^C$$

where X^C and Y^C are the complementary events of X and Y, respectively

Mutual and Independent Events

- **Independence (two events) -** Two events are independent if any one of the following equivalent statements is true

$$
\begin{aligned}
&(1) \quad P(A|B) = P(A) \\
&(2) \quad P(B|A) = P(B) \\
&(3) \quad P(A \cap B) = P(A)P(B)
\end{aligned}
$$

$$P(A' \cap B') = P(A')P(B')$$

- **Mutual Exclusive Events :** $E_1 \cap E_2 = \varnothing$

$$(E')' = E$$

Axioms of Probability

1. For any event A, the probability of the complementary event, written A^C, is given by $P(A) = 1 - P(A^C)$

2. The probability of an empty or impossible event, ϕ, is zero:

$$P(\phi) = 0$$

3. If event A is a subset of B, then – $P(A) \leq P(B)$

4. The probability that either events A or B occur or both occur is given by $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

5.If A and B are mutually exclusive events

$$P(A \cup B) = P(A) + P(B) \qquad P(A \cap B) = 0$$

6. If A1, A2, …, An are n events that form a partition of sample space S, then their probabilities must add up to 1:

$$P(A_1) + P(A_2) + \cdots + P(A_n) = \sum_{i=1}^{n} P(A_i) = 1$$

Joint Probability

- Let A and B be two events in a sample space. Then the joint probability of the two events, written as $P(A \cap B)$, is given by

$$P(A \cap B) = \frac{\text{Number of observations in } A \cap B}{\text{Total number of observations}}$$

Example

At an e-commerce customer service centre a total of 112 complaints were received. 78 customers complained about late delivery of the items and 40 complained about poor product quality.

(a) Calculate the probability that a customer will complain about both late delivery and product quality.

(b) What is the probability that a complaint is only about poor quality of the product?

Solution to Example

- Let A = Late delivery and B = Poor quality of the product. Let $n(A)$ and $n(B)$ be the number of events in favour of A and B. So $n(A) = 78$ and $n(B) = 40$. Since the total number of complaints is 112, hence

$$n(A \cap B) = 118 - 112 = 6$$

- Probability of a complaint about both delivery and poor product quality is

$$P(A \cap B) = \frac{n(A \cap B)}{\text{Total number of complaints}} = \frac{6}{112} = 0.0535$$

- Probability that the complaint is only about poor quality = 1 - $P(A)$ =

$$1 - \frac{78}{112} = 0.3035$$

- **Marginal probability** is simply a probability of an event *X*, denoted by *P*(*X*), without any conditions
- **Independent Events** : occurrence of one event (say event A) does not affect the probability of occurrence of the other event (event B).

$$P(A \cap B) = P(A) \times P(B).$$

- **Conditional Probability**: If A and B are events in a sample space, then the conditional probability of the event B given that the event A has already occurred, denoted by P(B|A), is defined as

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0$$

Bayes Theorem

- Bayes theorem is one of the most important concepts in analytics since several problems are solved using Bayesian statistics

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

- Using the two equations, we can show that

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

Terminologies in Bayes Theorem

1. *P*(*B*) is called the *prior probability* (estimate of the probability without any additional information).

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

2. *P*(*B*|*A*) is called the *posterior probability* (that is, given that the event *A* has occurred, what is the probability of occurrence of event *B*). That is, *post* the additional information (or additional evidence) that *A* has occurred, what is estimated probability of occurrence of *B*.

3. P(A|B) is called the likelihood of observing evidence A if B is true.

4. P(A) is the prior probability of A

Example

Black boxes used in aircrafts manufactured by three companies *A, B* and *C*. 75% are manufactured by *A*, 15% by *B*, and 10% by *C*. The defect rates of black boxes manufactured by *A, B*, and *C* are 4%, 6%, and 8%, respectively. If a black box tested randomly is found to be defective, what is the probability that it is manufactured by company *A*?

Solution to Example

- Let $P(A), P(B), P(C)$ be events corresponding to the black box being manufactured by companies A, B, and C, respectively, and $P(D)$ be the probability of defective black box. We are interested in calculating the probability $P(A|D)$.

$$P(A \mid D) = \frac{P(D \mid A) \times P(A)}{P(D)}$$

- Now $P(D|A) = 0.04$ and $P(A) = 0.75$. Using Eq.

$$P(D) = 0.75 \times 0.04 + 0.15 \times 0.06 + 0.10 \times 0.08 = 0.047$$

$$P(A \mid D) = \frac{0.04 \times 0.75}{0.047} = 0.6382$$

Example

Consider the scenario of selecting 100 fruits in which 45 are Orange and remaining are Apple. The fruits to be picked from from RED and BLUE buckets. The number of fruits in RED bucket are 40 in which 30 are orange and number of fruits in BLUE bucket are 60 in which 15 are orange. Find the following by applying Bayes Theorem

a.) Find all the probabilities

b) If we pick a fruit at random, what is the probability that it came out of the blue bucket?

c.) If we pick a fruit at random and it was turns out to be an orange, what is the probability that it came out of the blue bucket?

F: Fruit, B:Bucket, O:Orange, a=apple, r:red, b=blue

	F=o	**F=a**	
B=r	30	10	40
B=b	15	45	60
	45	55	100

Suppose a person goes to a cancer diagnosis centre for test and the test is +ve. The percentage of people with cancer test positive is 99% and percentage of people without cancer test negative is 99%. The percentage of population has cancer is 0.5%. Determine the following

a) What are the chances the person actually has cancer?
b) b) What are the random variables ?

Solution a and b together : random variables ,

State of desease D = {C, NC}

C - > Cancer Exists

NC - > Cancer does not exist (No cancer)

Result of the Test T = { + , - }

+ -> Test is positive

- -> Test I sNegative

Given Data : $P(+|C) = 0.99$

$P(-|C) = 0.01$

$P(-|NC) = 0.99$

$P(+|NC) = 0.01$

$P(C) = 0.05$ (Prior Probability)

Determine $P(C|+)$?

$P(C|+) = P(+|C)\ P(C)\ /\ P(+)$ (Apply bayes Theorom)

$P(+) = P(+|C)P(C)+P(+|NC)P(NC)$

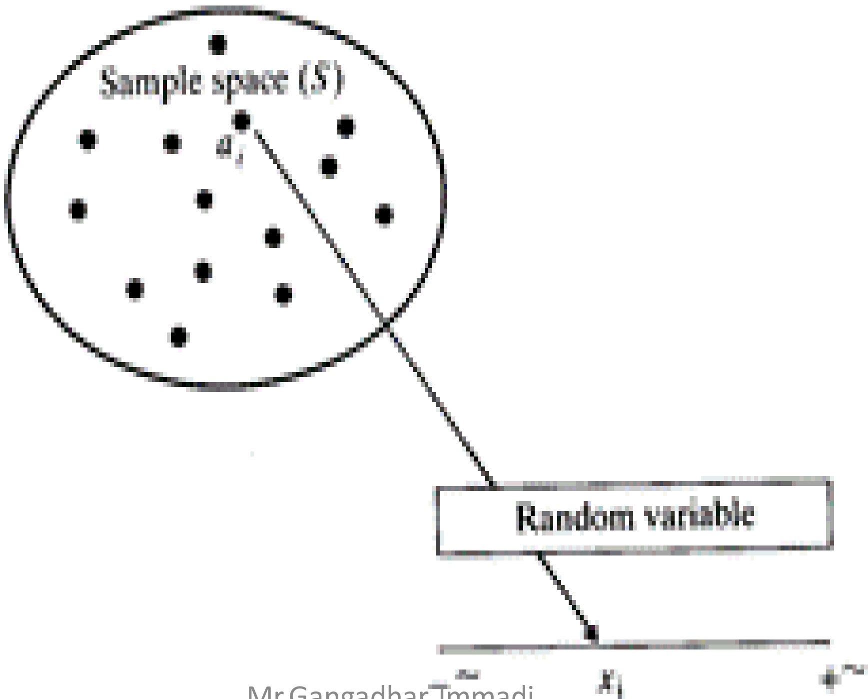
$= 0.99 \text{ X } 0.05 + 0.001 \text{ X } 0.995 = 0.005945$

$P(C|+) = 0.99 \text{ X } 0.005 / 0.005945 = 0.33$

So, chances the person actually has cancer is **0.833**

Random Variables

- A function that maps every outcome in the sample space to a real number.
- Robust and convenient way of representing the outcome of a random experiment



Discrete Random Variables

- If the random variable *X* can assume only a finite or countably infinite set of values, then it is called a discrete random variable.
- Examples of discrete random variables are:
  - Credit rating (usually classified into different categories such as low, medium and high or using labels such as AAA, AA, A, BBB, etc.).
  - Number of orders received at an e-commerce retailer which can be countably infinite.
  - Customer churn (the random variables take binary values, 1. Churn and 2. Do not churn).
  - Fraud (the random variables take binary values, 1. Fraudulent transaction and 2. Genuine transaction).
  - Any experiment that involves counting (for example, number of returns in a day from customers of e-commerce portals such as Amazon, Flipkart; number of customers not accepting job offers from an organization).

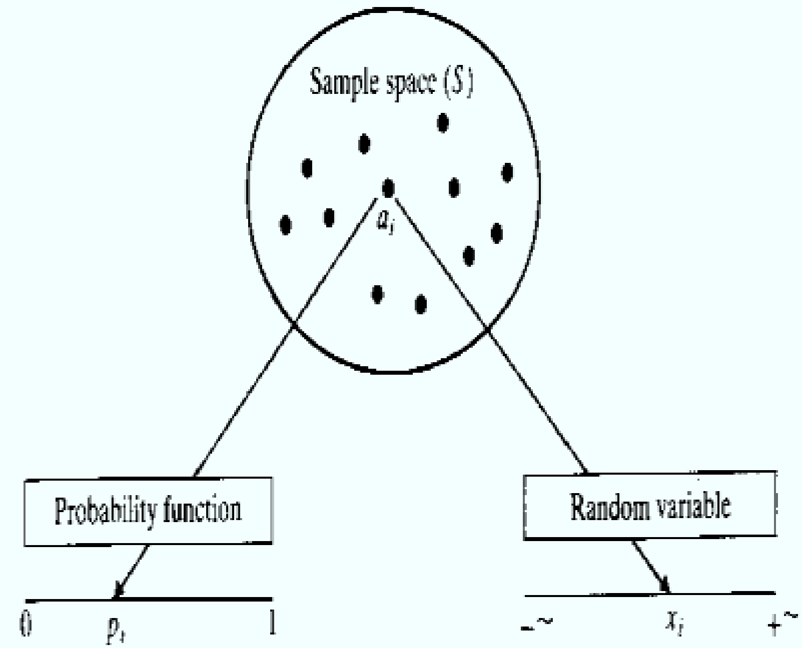
Continuous Random Variables

- A random variable *X* which can take a value from an infinite set of values is called a continuous random variable
- Examples of continuous random variables are listed below:
  - Market share of a company (which take any value from an infinite set of values between 0 and 100%).
  - Percentage of attrition among employees of an organization.
  - Time to failure of engineering systems.
  - Time taken to complete an order placed at an e-commerce portal.
  - Time taken to resolve a customer complaint at call and service centers.

Probability mass function

- For a discrete random variable, the probability that a random variable X taking a specific value x_i, $P(X = x_i)$, is called the probability mass function $P(x_i)$.

- That is, a probability mass function is a function that maps each outcome of a random experiment to a probability



X	x_1	x_2	…………..	x_n
P	p_1	p_2	…………..	p_n

Probability distribution of a random variable

Expected Value

- Expected value (or mean) of a discrete random variable is given by

$$E(X) = \sum_{i=1}^{n} x_i P(x_i)$$

- Where x_i is the specific value taken by a discrete random variable X and $P(x_i)$ is the corresponding probability, that is, $P(X = x_i)$.

Variance and Standard Deviation

Variance of a discrete random variable is given by

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \left[x_i - E(X) \right]^2 \times P(x_i)$$

Standard deviation of a discrete random variable is given by

$$\sigma = \sqrt{VAR(X)}$$

Probability Density Function (pdf)

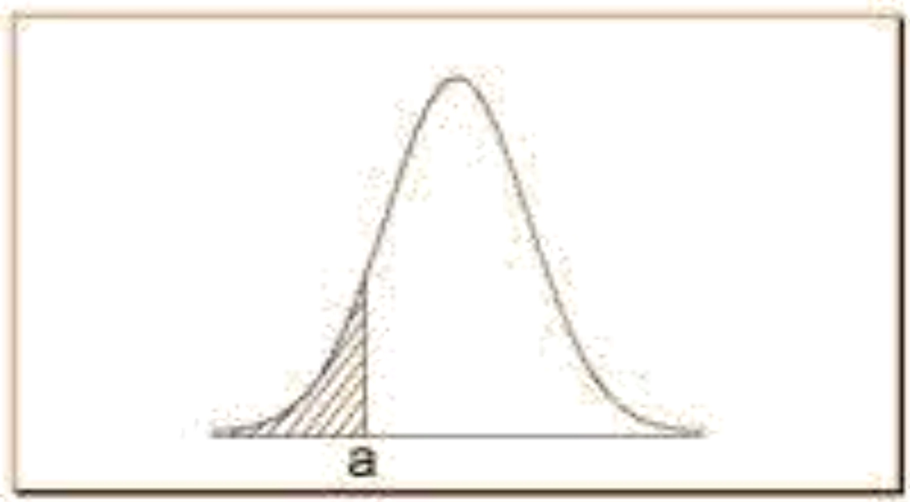
- The probability density function, $f(x_i)$, is defined as probability that the value of random variable X lies between an infinitesimally small interval defined by x_i and $x_i + \delta x$

$$f(x) = \lim_{\delta x \to 0} \frac{P(x_i \le X \le x_i + \delta x)}{\delta x}$$

Cumulative Distribution Function (CDF)

- The cumulative distribution function (CDF) of a continuous random variable is defined by

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$



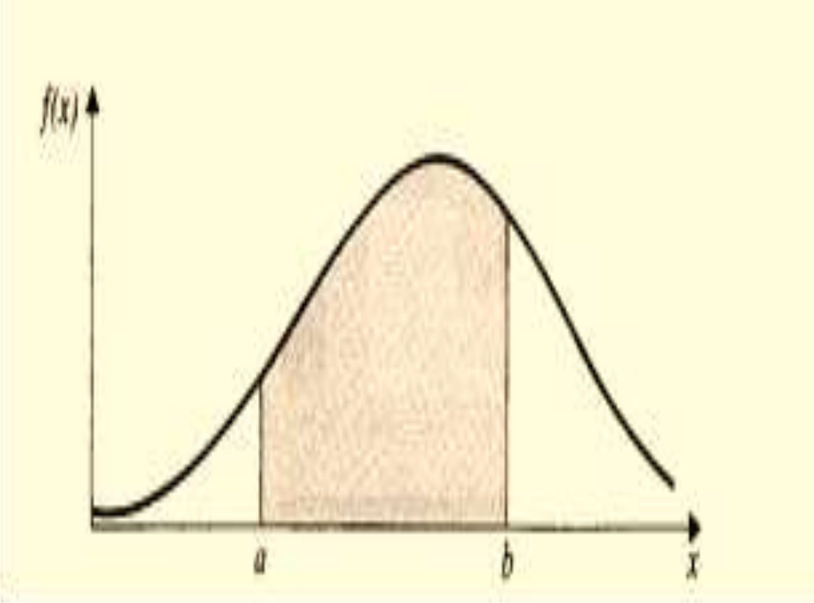
Cumulative distribution function

- Probability density function and cumulative distribution function of a continuous random variable satisfy the following properties
- $f(x) \geq 0$

$$F(\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$$

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$$

The probability between two values a and b, $P(a \leq X \leq b)$, is the area between the values a and b under the probability density function



- The expected value of a continuous random variable, $E(X)$, is given by

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

- The variance of a continuous random variable, Var(X), is given by

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} [x - E(x)]^2 f(x)dx$$

Binomial Distribution

- A random variable X is said to follow a Binomial distribution when
  - The random variable can have only two outcomes *success* and *failure* (also known as Bernoulli trials).
  - The objective is to find the probability of getting k successes out of n trials.
  - The probability of success is p and thus the probability of failure is $(1 - p)$.
  - The probability p is constant and does not change between trials.

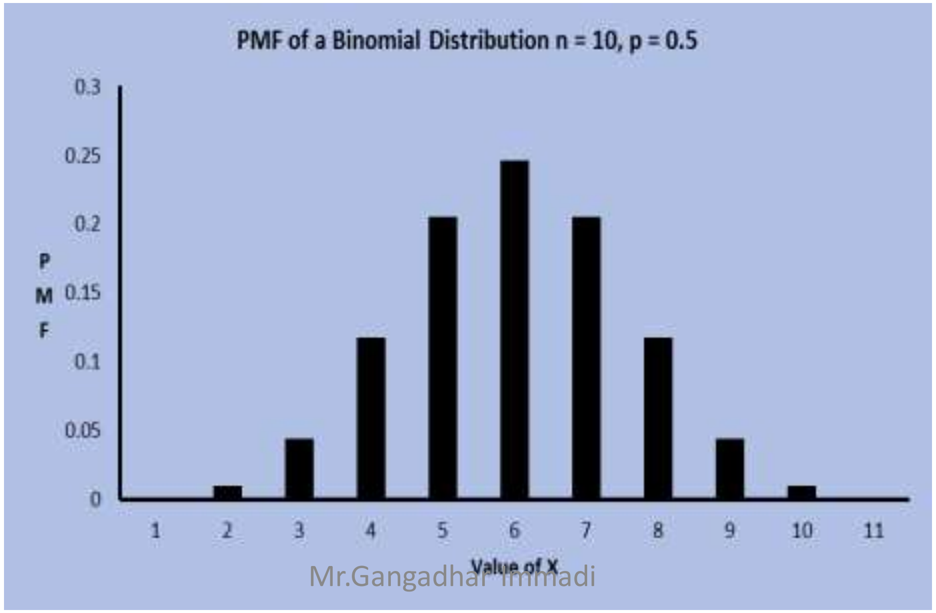
Probability Mass Function (PMF) of Binomial Distribution

- The PMF of the Binomial distribution (probability that the number of success will be exactly *x* out of *n* trials) is given by

$$PMF(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad 0 \le x \le n$$

Where

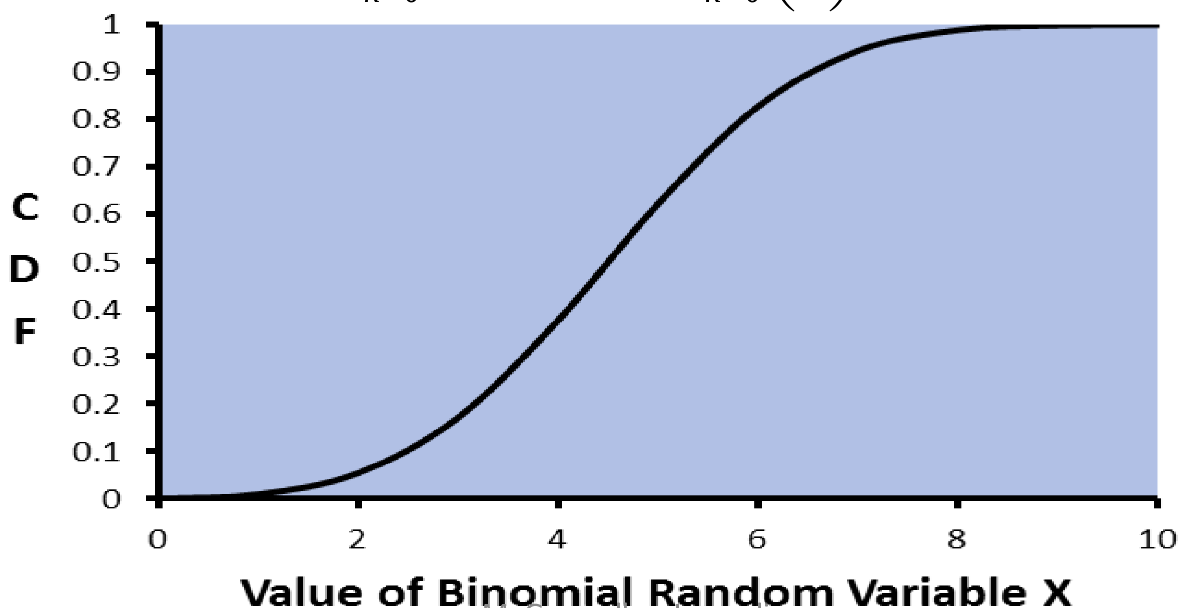
$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$



Cumulative Distribution Function (CDF) of Binomial Distribution

- CDF of a binomial distribution function, F(x), representing the probability that the random variable *X* takes value less than or equal to *a*, is given by
-

F(a) = $$P(X \le a) = \sum_{k=0}^{a} P(X = k) = \sum_{k=0}^{a} \binom{n}{k} p^k (1-p)^{n-k}$$



Mean and Variance of Binomial Distribution

The Mean of a binomial distribution is given by:

$$\text{Mean} = E(X) = \sum_{x=0}^{n} x \times \text{PMF}(x) = \sum_{x=0}^{n} x \times \binom{n}{x} p^x (1-p)^{n-x} = np$$

The variance of a binomial distribution is given by

$$\text{Var}(X) = \sum_{x=0}^{n} (x - E(X))^2 \times \text{PMF}(x) = \sum_{x=0}^{n} (x - E(X))^2 \times \binom{n}{x} p^x (1-p)^{n-x} = np(1-p)$$

If the number of trials (n) in a binomial distribution is large, then it can be approximated by normal distribution with mean np and variance npq.

Example

Fashion Trends Online (FTO) is an e-commerce company that sells women apparel. It is observed that about 10% of their customers return the items purchased by them for many reasons (such as size, color, and material mismatch). On a particular day, 20 customers purchased items from FTO. Calculate:

(a) Probability that exactly 5 customers will return the items.

(b) Probability that a maximum of 5 customers will return the items.

(c) Probability that more than 5 customers will return the items purchased by them.

(d) Average number of customers who are likely to return the items.

(e) The variance and the standard deviation of the number of returns.

Solution

In this case, the value of n = 20 and p = 0.1.

(a) Probability that exactly 5 customers will return the items purchased is

$$P(X=5)=\binom{20}{5}\times(0.1)^5\times(0.9)^{15}=0.03192$$

(b) Probability that a maximum of 5 customers will return the items purchased is

$$P(X\le 5)=\sum_{k=0}^{5}\binom{20}{k}\times(0.1)^k\times(0.9)^{20-k}=0.9887$$

(c) Probability that more than 5 customers will return the product is

$$P(X>5)=1-P(X\le 5)=1-\sum_{k=0}^{5}\binom{20}{k}\times(0.1)^k\times(0.9)^{20-k}=1-0.9887=0.0113$$

(d) The average number of customers who are likely to return the items is

$$E(X) = n \times p = 20 \times 0.1 = 2$$

(e) Variance of a binomial distribution is given by

$$Var(X) = n \times p \times (1-p) = 20 \times 0.1 \times 0.9 = 1.8$$

and the corresponding standard deviation is 1.3416

Example

Die another Day hospital recruits nurses frequently to manage high attrition among the nursing saff. Not all job offers from DAD hospital are accepted. Based on the past recruitment data only 70% of of offers rolled out by DAD hospital are accepted.

(a) If 10 offers are made, what is the probability that more thann 5 and less than 8 candidates will accept the offer from DAD Hospital?

(b) During March 2017, DAD required 14 new nurses to manage attrition. What should be the number of offers made by DAD hospital so that average numbers of nurses accepting the offer is 14?

Poisson Distribution

- Poisson distribution is used when we have to find the probability of number of events
- The probability mass function of a Poisson distribution is given by

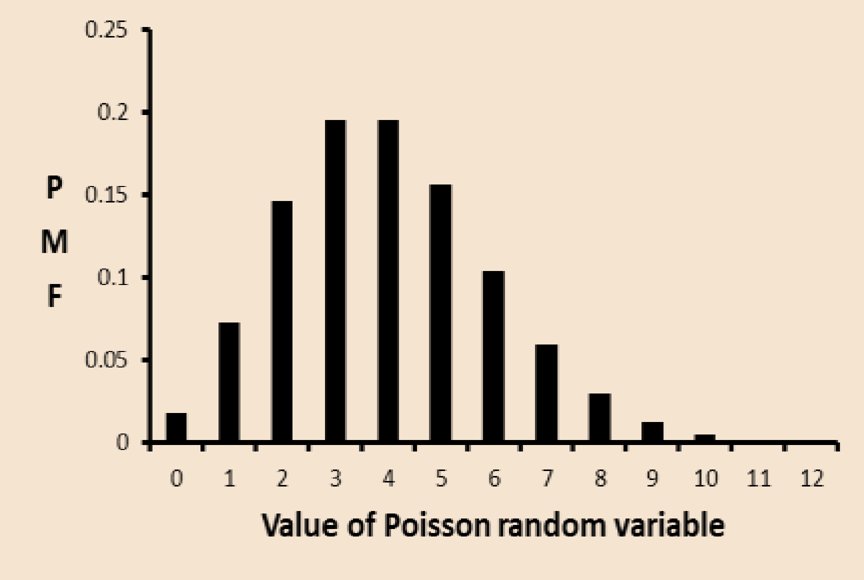
$$P(X = k) = \frac{e^{-\lambda} \times \lambda^{k}}{k!}, \quad k = 0, 1, 2, \ldots$$

- where λ is the rate of occurrence of the events per unit of measurement
- Cumulative distribution function of a Poisson distribution is given by

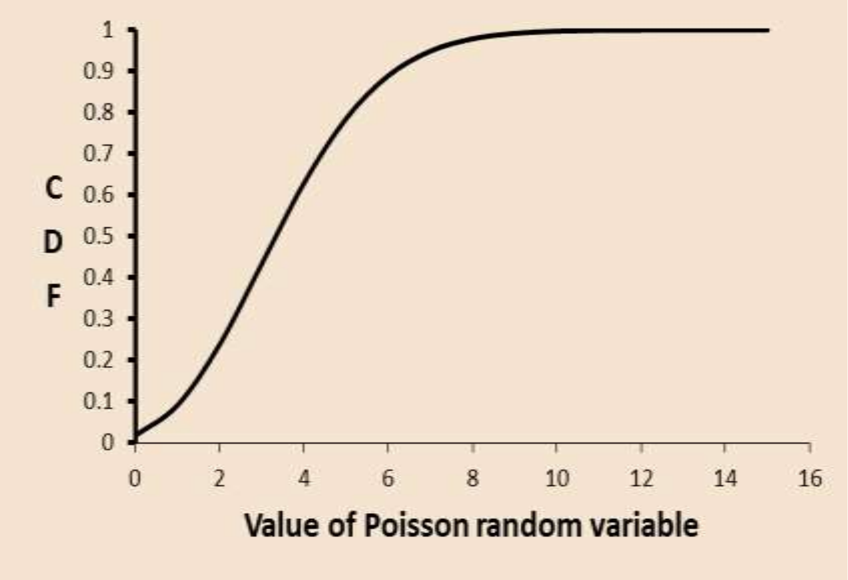
$$P[X \leq k] = \sum_{i=0}^{k} \frac{e^{-\lambda} \times \lambda^{k}}{k!}$$

- The mean and variance of a Poisson random variable are given by $E(X) = \lambda$ and $\mathrm{Var}(X) = \lambda$

Probability mass function of a Poisson random variable (λ = 4).



Cumulative distribution function of a Poisson random variable (λ = 4).



Example

On average, about 20 customers per day cancel their order placed at Fashion Trends Online. Calculate the probability that the number of cancellations on a day is exactly 20 and the probability that the maximum number of cancellations is 25

Solution

The probability that the number of cancellations is exactly 20 is given by

$$P(X = 20) = \frac{e^{-20} 20^{20}}{20!} = 0.0888$$

Probability that the maximum number of cancellation will be 25 is given by

$$P(X \leq 25) = \sum_{k=0}^{25} \frac{e^{-20} 20^{k}}{k!} = 0.8878$$

Normal Distribution

- Normal distribution, also known as **Gaussian distribution**, is one of the most popular continuous distribution in the field of analytics especially due to its use in multiple contexts
- The probability density function and the cumulative distribution function are given by

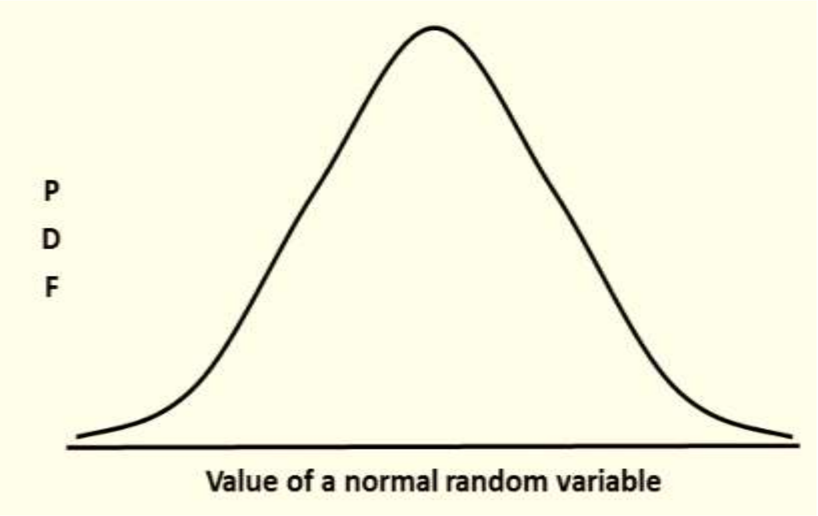
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < +\infty$$

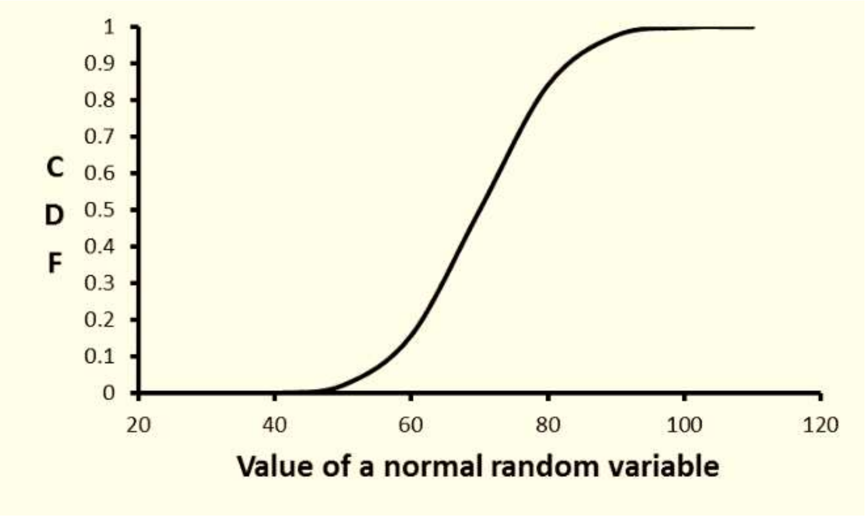
$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt, \qquad -\infty < x < +\infty$$

- Here μ and σ are the mean and standard deviation of the normal distribution

Probability density function of a normal distribution

Cumulative distribution function of a normal distribution.





Properties of Normal Distribution

1. Theoretical normal density functions are defined between $-\infty$ and $+\infty$.

2. It is a two parameter distribution, where the parameter μ is the mean (location parameter) and the parameter σ is the standard deviation (scale parameter).

3. All normal distributions have symmetrical bell shape around mean μ (thus it is also median). μ is also the mode of the normal distribution, that is, μ is the mean, median as well as the mode.

4. For any normal distribution, the areas between specific values measured in terms of μ and σ are given by:

Value of Random Variable	Area under the Normal Distribution (CDF)
$\mu - \sigma \leq X \leq \mu + \sigma$ (area between one sigma from the mean)	0.6828
$\mu - 2\sigma \leq X \leq \mu + 2\sigma$ (area between two sigma from the mean)	0.9545
$\mu - 3\sigma \leq X \leq \mu + 3\sigma$ (area between three sigma from the mean)	0.9973

5. Any linear transformation of a normal random variable is also normal random variable. That is, if X is a normal random variable, then the linear transformation AX + B (where A and B are two constants) is also a normal random variable.

- If X_1 and X_2 are two independent normal random variables with mean μ_1 and μ_2 and variance σ_1^2 and σ_2^2 respectively, then X1 + X2 is also a normal distribution with mean μ1 + μ2 and variance $\sigma_1^2 + \sigma_2^2$

- Sampling distribution of mean values a large sample drawn form a population of any distribution is likely to follow a normal distribution, this result is known as the central limit theorem

Standard Normal Variable

- A normal random variable with mean $\mu = 0$ and $\sigma = 1$ is called the standard normal variable and usually represented by Z
- The probability density function and cumulative distribution function of a standard normal variable are given by

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

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dz$$

- By using the following transformation, any normal random variable *X* can be converted into a standard normal variable

$$Z = \frac{X - \mu}{\sigma}$$

- The random variable X can be written in the form of a standard normal random variable using the relationship

$$X = \mu + \sigma Z$$

Chi-Square Distribution

- Chi-square distribution with k degrees of freedom [denoted as $\chi^2(k)$ distribution] is a non-parametric distribution which is obtained by adding square of k independent standard normal random variables.
- Consider a normal random variable X_1 with mean μ_1 and standard deviation σ_1. Then we can define Z_1 (the standard normal random variable) as

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1}$$

- Then,

$$Z_1^2 = \left(\frac{X_1 - \mu_1}{\sigma_1} \right)^2$$

is a chi-square distribution with one degree of freedom [χ2(1)]

- Let X_2 be a normal random variable with mean μ_2 and standard deviation σ_2 and Z_2 is the corresponding standard normal variable. Then the random variable given by $Z_1^2 + Z_2^2$

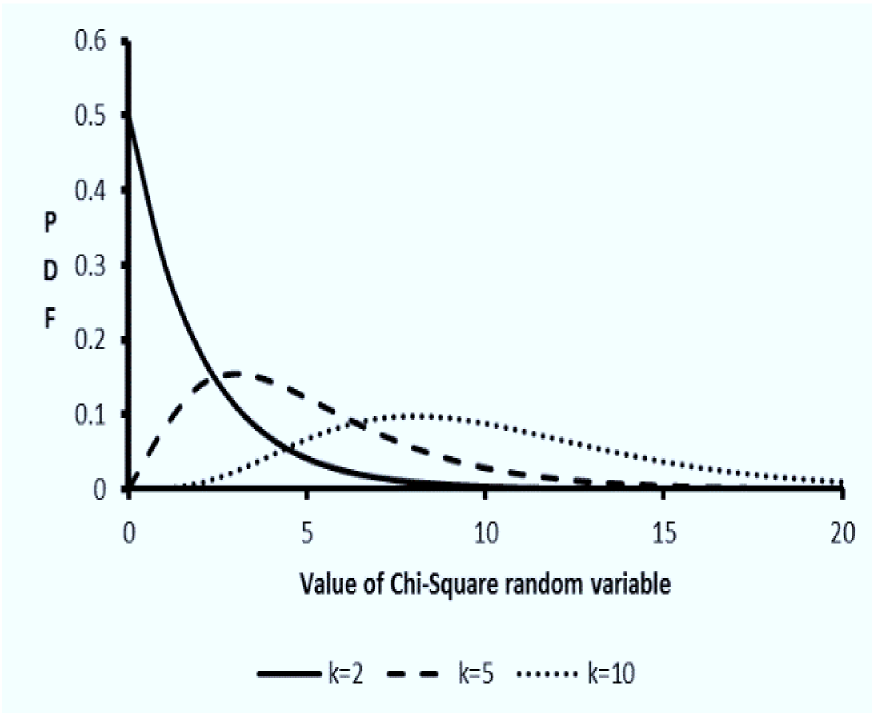
$$Z_1^2 + Z_2^2 = \left(\frac{X_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{X_2 - \mu_2}{\sigma_2}\right)^2$$

is a chi-square distribution with 2 degrees of freedom.

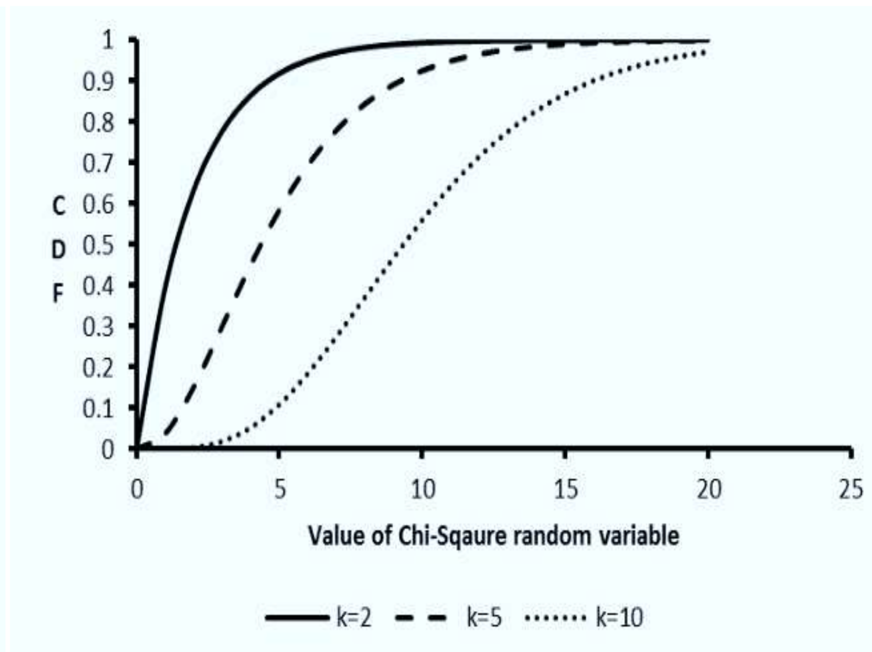
- A chi-square distribution with k degrees of freedom is given by sum of squares of standard normal random variables Z1, Z2, …, Zk obtained by transforming normal random variables X1, X2, …, Xk with mean values μ1, μ2, …, μk and corresponding standard deviations σ1, σ2, …, σk. That is

$$\chi^2(k) = Z_1^2 + Z_2^2 + \ldots + Z_k^2 = \left(\frac{X_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{X_2 - \mu_2}{\sigma_2}\right)^2 + \ldots + \left(\frac{X_k - \mu_k}{\sigma_k}\right)^2$$

- Probability density function of chi-square distribution for different values of *k*



- Cumulative distribution of chi-square distribution with k degrees of freedom



Properties of chi-square distribution

- The mean and standard deviation of a chi-square distribution are k and $\sqrt{2k}$ where k is the degrees of freedom

- As the degrees of freedom k increases the probability density function of a chi-square distribution approaches normal distribution.

- Chi-square goodness of fit test is one of the popular tests for checking whether a data follows a specific probability distribution.

Student's *t*-Distribution

- Student's *t*-distribution (or simply *t*-distribution) arises while estimating the population mean of a normal distribution using sample which is either small and/or the population standard deviation is unknown

- Assume that $X_1, X_2, \ldots, X_n$ are n observations (that is, sample of size n) from a normal distribution with mean μ and standard deviation σ. Let

$$\bar{X} = \sum_{i=1}^{n} X_i$$

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

- where $\bar{X}$ and S are mean and standard deviation estimated from the sample X1, X2, …, Xn. Then the random variable t defined by

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows a t-distribution with (n – 1) degrees of freedom.

Properties of t-distribution:

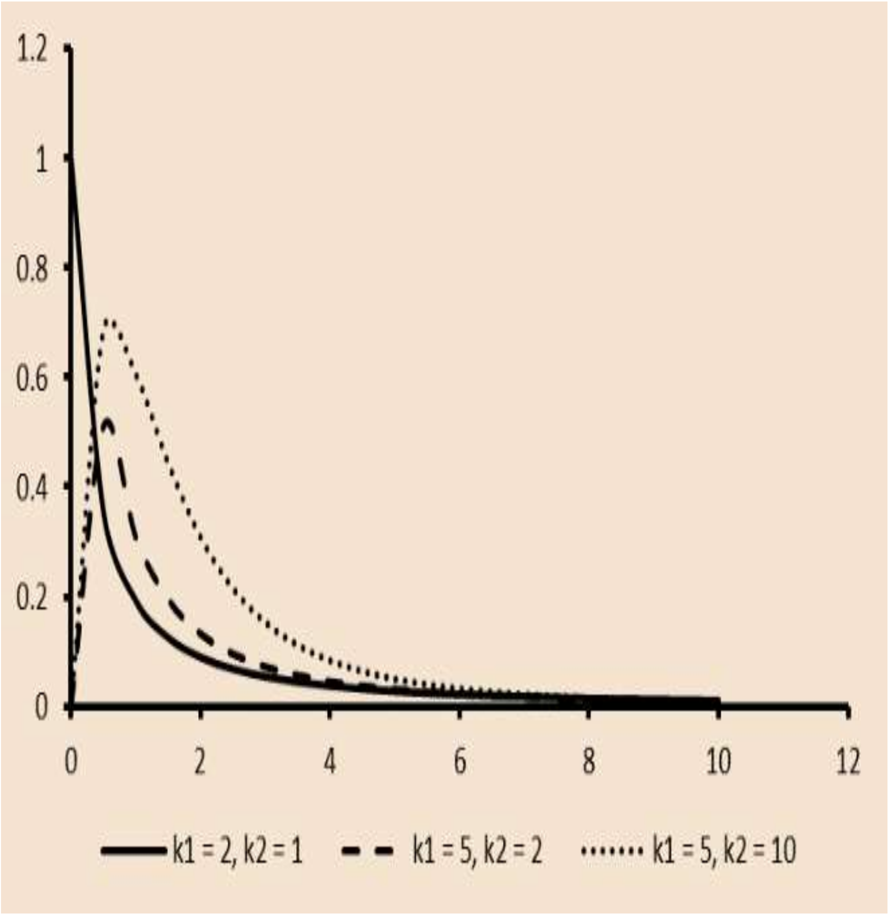
- The mean of a t distribution with 2 or more degrees of freedom is 0.
- The standard deviation of t-distribution is $\frac{n}{n-2}$ for n > 2, where n is the number of degrees of freedom.
- As the degrees of freedom n increases the probability density function of a t-distribution approaches the density function of standard normal distribution. For n > 120, the difference between the area under probability density function of a t-distribution is very close to the area under a standard normal distribution.
- t-distribution is an important distribution for hypothesis testing of means of a population and for comparing means of two populations.

F-Distribution

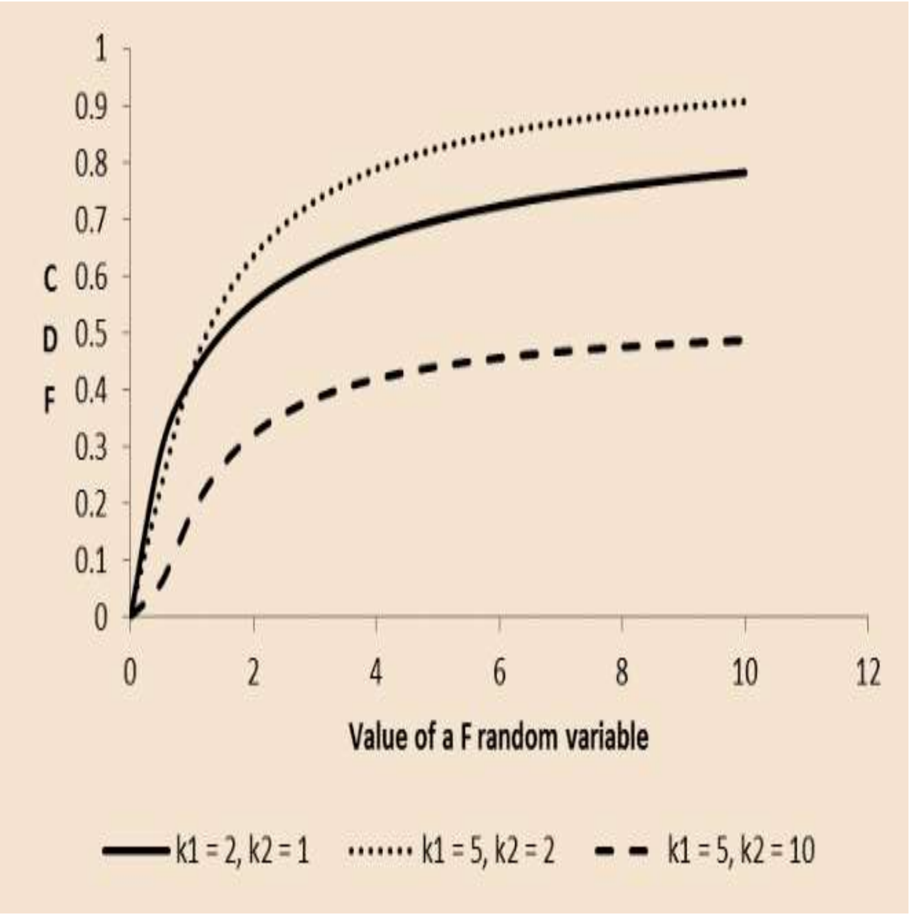
F-distribution (short form of Fisher's distribution named after statistician Ronald Fisher) is a ratio of two chi-square distributions. Let Y_1 and Y_2 be two independent chi-square distributions with k_1 and k_2 degrees of freedom, respectively. Then the random variable X is defined as

$$X = \frac{Y_1 / k_1}{Y_2 / k_2}$$

- Probability density function of *F*-distribution



- Cumulative density function of F-distribution



Properties of F distribution:

- Mean of F-distribution is $\frac{k_2}{k_2 - 2}$, for k2 > 2.
- Standard deviation of F-distribution is $\sqrt{\frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}}$ for k2 > 4.
- F-distribution is non-symmetrical and the shape of the distribution depends on the values of k1 and k2.
- F-distribution is used in Analysis of Variance to test the mean values of multiple groups