Probability Theory and Random Variables

Dhruva Hegde

1 Basics of Probability Theory

Probability is a measure of uncertainty of occurrence of a particular event. Probability in it's most primitive form is defined as the ratio of number of favourable outcomes and the total number of possible outcomes.
$\implies P(E) = p/q$
where E is the event of interest, P(E) is the probability of occurrence of the event E, p is the outcomes favourable for E, q is total possible outcomes.

1.1 Sample Space

The process used to collect data and measure probability is called an experiment. And experiment that does not produce the same outcome every time is called **Random Experiment**.

The **Sample Space**, Ω is defined as the set of all possible outcomes of a random experiment.

The sample space set should be

- Mutually exclusive, meaning occurrence of any outcome should mean the non-occurrence of other outcomes.
- Collective exhaustive, meaning there should be no other possible outcome.

- Of the right granularity, meaning it shouldn't include unnecessary events that are not relevant to the experiment.

Types of Sample Space

1. Discrete Sample Space
   The outcomes are countable. For example, rolling a dice has 6 possible outcomes.

2. Continuous Sample Space
   The outcomes are uncountable. For example, throwing a dart on a dart board, it can land on any point on the board.

Note that there might be sample spaces that are discrete but infinite as in, a countably infinite sample space.

1.2 Probability Axioms

Probability is assigned to different events of a sample space.
The basic axioms are probability are

- Non-negativity: $P(E) \geq 0$

- Normalization: $P(\Omega) = 1$

- Additivity: if there are no common outcomes under events A and B i.e $P(A \cap B) = 0$, then $P(A \cup B) = P(A) + P(B)$

Discrete uniform law
The probability of an event in a discrete sample space is defined as,
$\boxed{P(E) = n(E)/n(\Omega)}$ where $n(E)$ is number of outcomes under event A and $n(\Omega)$ is number of outcomes in the sample space.

Example: The probability of getting a number greater than 4 in a roll of a die. $n(\Omega) = 6 \ \ [1, 2, 3, 4, 5, 6]$, $n(E) = 3 \ \ [4, 5, 6]$, $\implies P(E) = 3/6 = 0.5$

Example for countable infinite sample space:
Consider tossing a fair coin till first heads is obtained. In this setup, the coin is repeatedly tossed till a heads occurs and number of tosses it takes is noted. The number of tosses required can be anything from 1 to infinite.

Here, probability is generally calculated using sum of infinite series. To find probability the that heads occurs in an odd index, add all odd index probabilities.
P(Odd) = P(1) + P(3) + P(5) + ... = 1/2 + 1/8 + 1/32 + ... = 2/3.

Continuous uniform law
The probability of an event in a continuous sample space is defined using area. Meaning, the area under the entire sample space is taken as 1 and the area for the required range (of favourable outcomes) is calculated.

Note that the probability of occurrence of any one specific outcome (atomic event) in a continuous sample space is 0 because area of a point is 0.

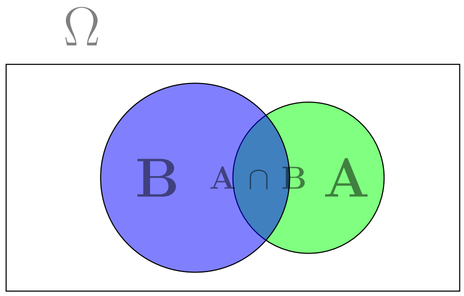
Example: The probability of the dart landing exactly on the bulls-eye of a dart board is 0. The probability of the dart landing within $\frac{1}{10}^{th}$ the radius of the dart board is 0.01.

1.3 Conditional Probability

Conditional probability is defined as the probability of occurrence of an event, given that another event in the sample space has already occurred.
$P(A|B)$ is the probability of occurrence of A, given B has occurred.
To calculate $P(A|B)$, since it is known that B has already occurred, the sample space Ω is reduced to just B and the probability of occurrence of A in the new sample space B is calculated.

$$\boxed{\therefore P(A|B) = \frac{P(A \cap B)}{P(B)}; \quad P(B) \neq 0}$$

It should also be noted that if the occurrence of A means the non-occurrence of B, it means A and B are dependent.



1.3.1 Multiplication Rule

Conditioning on 2 events -

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

Conditioning on 3 events -

$$P(A \cap B \cap C) = P(A).P(B|A).P(C|A \cap B)$$

Can be extended to multiple events in similar manner.

1.3.2 Total Probability Theorem

If the sample space is partitioned into $A_1,\ A_2,\ A_3, ...$ and B is an event in the sample space where conditional probabilities of B with respect to $A_1,\ A_2,\ A_3, ...$ are known, then to find probability of B,

$$\boxed{P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) + ...}$$

1.3.3 Bayes' Theorem

If the sample space is partitioned into $A_1,\ A_2,\ A_3, ...$ and B is an event in the sample space where conditional probabilities of B with respect to $A_1,\ A_2,\ A_3, ...$ are known, then to find conditional probability of A_i with respect to B,

$$\boxed{P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\Sigma P(A_i)P(B|A_i)} = \frac{P(A_i)P(B|A_i)}{P(B)}}$$

1.4 Independent Events

Two events A and B are said to be independent if the occurrence of one of them does not affect the occurrence of another.
Meaning, probability of occurrence of A and conditional probability of occurrence of A, given B has occurred will be the same (and vice versa).
$P(A|B) = P(A);\quad P(B|A) = P(B)$
This implies for A and B to be independent, $P(A \cap B) = P(A)P(B)$.

Assuming $P(A) \neq 0$ and $P(B) \neq 0$,

- If $P(A \cap B) = 0$, it means the occurrence of A guarantees non-occurrence of B (and vice versa), therefore A and B are dependent.
- If $P(A \cap B) = 1$, it means the occurrence of A guarantees occurrence of B (and vice versa), therefore A and B are still dependent.

This definition of independence is valid for more than 2 events as well. A set of events are said to be mutually independent of each other if the probability of occurrence of intersection of all the events is equal to product of probabilities of each of the events taken individually.

1.4.1 Conditioning on independence

Two events A and B are conditionally independent after C has already occurred if $P(A \cap B|C) = P(A|C)P(B|C)$.
Note that independence of A and B in general does not imply conditional independence (and vice versa).

Meaning, $P(A \cap B) = P(A)P(B) \nRightarrow P(A \cap B|C) = P(A|C)P(B|C)$.

Pairwise independence
3 events A,B,C are said to have pairwise independence if they satisfy
$P(A \cap B) = P(A)P(B), P(B \cap C) = P(B)P(C)$ and $P(C \cap A) = P(C)P(A)$.
This definition can be extended for a set of more than 3 events.

Note that pairwise independence does not imply mutual independence of all 3 events. Meaning, if A,B,C are pairwise independent, it not need satisfy that $P(A \cap B \cap C) = P(A)P(B)P(C)$. Also, if this condition is satisfied, it does not imply pairwise independence.

(Conclusion, pairwise independence and conditional independence are independent of general/direct/mutual independence)

2 Random Variables

A random variable is a numerical description of the outcome of a statistical experiment. It is an assignment of a value (number) to every possible outcome of the experiment.

For example, for the event of tossing a fair coin, a random variable X can be defined as

$$X = \begin{cases} 0 & \text{if } Heads \\ 1 & \text{if } Tails \end{cases}$$

This is a discrete random variable, since the outcomes are countable values.

Consider the event of choosing a random animal from a zoo and finding it's weight. The random variable is defined as X = exact weight of the random animal. Here, it can take any value from minimum possible weight to maximum possible weight and the possible outcomes are not countable, hence it is a continuous random variable.

2.1 Discrete Random Variables

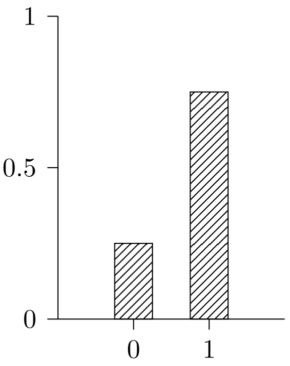
Discrete Random Variables are represented using Probability Mass Functions.
Probability Mass Function is a function that gives the probability that a discrete random variable is exactly equal to some value.
The PMF is represented as $p_X(x)$ where X represents the random variable and x is the value.
The notation denotes the probability that the random variable X takes on the value x i.e $\boxed{p_X(x) = P(X = x)}$.

Consider another example where a fair coin is tossed twice and a random variable X is defined as,

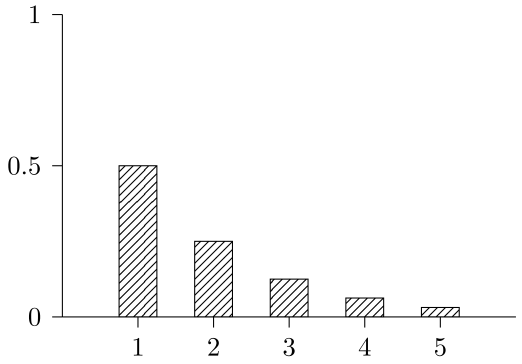
$$X = \begin{cases} 1 & \text{if Heads occurs} \\ 0 & \text{if Heads does not occur} \end{cases}$$

.
The 4 possible outcomes are $[HH\ , HT\ , TH\ , TT]$ and from the definition of X, P(X=0) = 0.25 and P(X=1) = 0.75 since head occurs in 3 of the 4 cases.

The corresponding PMF graph will be as follows-



Taking the example of countably infinite case i.e tossing a fair coin till heads is first obtained.
X is the random variable which is the number of tosses before first head occurs.
The probabilities assigned to the discrete values of the random variable are P(X=1) = 0.5, P(X=2) = 0.25, ...
The corresponding PMF will be an infinitely extending and decaying graph.



For a discrete random variable to be a valid random variable, it's PMF must satisfy the following-

- $P_X(x) \geq 0$, meaning each value must have probability greater than or equal to zero.
- $\sum_x P_X(x) = 1$, meaning sum of probabilities of all values must be equal to one.

2.1.1 Expectation

Expectation is the mean or average value of the random experiment.
It is given by,

$$\boxed{E(X) = \sum_x xP_X(x)}$$

Consider X to be a random variable and another random variable Y is defined as a function of X, i.e Y = g(X).
In this case, if $P_X(x)$ is the PMF of X and $P_Y(y)$ is the PMF of Y, then the expectation of Y is given by,

$$E(Y) = \sum_y yP_Y(y) = \sum_x g(x)P_X(x)$$

Properties of expectation of discrete random variable (where α and β are constants)

- $E(\alpha) = \alpha$
- $E(\alpha X) = \alpha E(X)$
- $E(\alpha X + \beta) = \alpha E(X) + \beta$

2.1.2 Variance

The n^th moment of any random variable X with PMF $P_X(x)$ is defined as $E(X^n) = \sum_x x^n P_X(x)$.
The variance of a random variable is it's second moment subtracted by the mean square value. Meaning, the variance of X is given by,
$var(X) = \sum_x x^2 P_X(x) - [\sum_x xP_X(x)]^2$

$$\therefore \boxed{var(X) = E[X^2] - (E[X])^2}$$

Properties of variance of discrete random variable (where α and β are constants)

- $var(\alpha) = 0$

- $var(\alpha X) = \alpha^2 var(X)$
- $var(\alpha X + \beta) = \alpha^2 var(X)$

Note that var(X) is always a positive quantity.

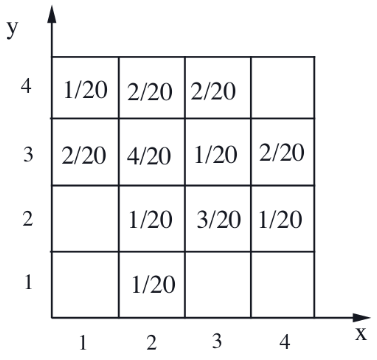
Standard deviation of X is defined as the square root of the variance of X and is denoted by σ_X. $\implies \sigma_x = \sqrt{var(X)}$

2.1.3 Joint PMF

2 (or more) different discrete random variables are sometimes necessary to quantify a certain experiment. In these cases, the joint probability mass function has to be used.
If X and Y are two random variables, then their joint PMF is given by,

$$P_{X,Y}(x, y) = P(X = x \textit{ and } Y = y)$$

A joint PMF is best represented as a table. A random example is shown below.



y					
4	1/20	2/20	2/20		
3	2/20	4/20	1/20	2/20	
2		1/20	3/20	1/20	
1		1/20			
	1	2	3	4	x

The common properties such as sum of all probabilities must be 1 and each entry must be greater than 0 will hold.

Individual PMF can also be found from joint PMF.
$P_X(x) = \sum_y P_{X,Y}(x, y)$ and $P_Y(y) = \sum_x P_{X,Y}(x, y)$

A few examples from the given joint PMF, $P_{X,Y}(1,3) = 2/20,\ P_X(2) = 6/20$ and $P_Y(1) = 1/20$.

Conditional Probabilities in Joint PMF
The joint PMF can be used to find conditional probabilities of the random variables. $P_{X|Y}(x|y) = P(X = x \mid Y = y)$ The above expression finds the conditional probability of occurrence of a certain value for X given that the value of Y is fixed at a certain value.

For example, from the given joint PMF, $P_{X|Y}(2|3) = P(X = 2|Y = 3) = P(X = 2 \cap Y = 3|Y = 3) = \frac{(4/20)}{9/20} = \frac{4}{9}$

The conditional PMF $P_{X|Y}(x|3)$ will consist of the probabilities [2/9, 4/9, 1/9, 2/9], hence they are scaled versions of the entries of Y = 3 in the joint PMF, such that the sum adds up to 1.

In a joint PMF, X and Y are independent discrete random variables if $P_{X,Y}(x,y) = P(X = x)P(Y = y)$.

Note that in the given PMF, X and Y are not independent. However, considering a conditional universe where $X \leq 2$ and $Y \geq 3$, it can be observed that in the new conditional PMF, X and Y are independent. This is further indication that conditioning may affect independence.

Expectation and Variance of Joint PMF
In general, expectation is a linear operation, meaning it satisfies
E(X+Y) = E(X) + E(Y).

If X and Y are independent, then E(XY) = E(X)E(Y).

If g(X) and h(Y) are functions of X and Y (where X and Y are independent), it means that g(X) and h(Y) are also independent, and hence
E[g(X)h(Y)] = E[g(X)]E[h(Y)].

Also, in case of independence of X and Y, the variances satisfy
var(X + Y) = var(X) + var(Y).
Note that if Z = X - kY, then var(Z) = var(X) + k^2var(Y).

2.2 Continuous Random Variables

Continuous Random Variables are represented using Probability Density Functions.
Probability Density Function is a function that specifies the probability of the random variable falling within a particular range of values (as opposed to taking on any one value).
The PDF is represented as $f_X(x)$ where X represents the random variable and x is the dummy variable.
The integration of the PDF between the intervals 'a' and 'b' denotes the probability that the random variable X takes a value between the intervals.

$$\boxed{P(a \leq X \leq b) = \int_a^b f_X(x)dx}$$

The important property of any valid PDF is, the area under the curve should be equal to 1.
$\implies P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f_X(x)dx = 1$

2.2.1 Mean and Variance

Concept of mean in PDF is same as expectation in PMF, it uses integration instead of summation. $E(X) = \int_{-\infty}^{\infty} xf_X(x)dx$

Similarly the variance. $var(X) = \int_{-\infty}^{\infty}(x - E(x))^2 f_X(x)dx$

2.2.2 Cumulative Distribution Functions

The CDF is actually defined for both continuous and discrete random variables. In general, it is denoted by $F_X(x)$ and is defined as,

$$\boxed{F_X(x) = P(X \leq x)}$$

The CDF graph is the accumulation of probabilities of all events upto the value x.
For continuous random variables given by PDF $f_X(x)$, the CDF will be

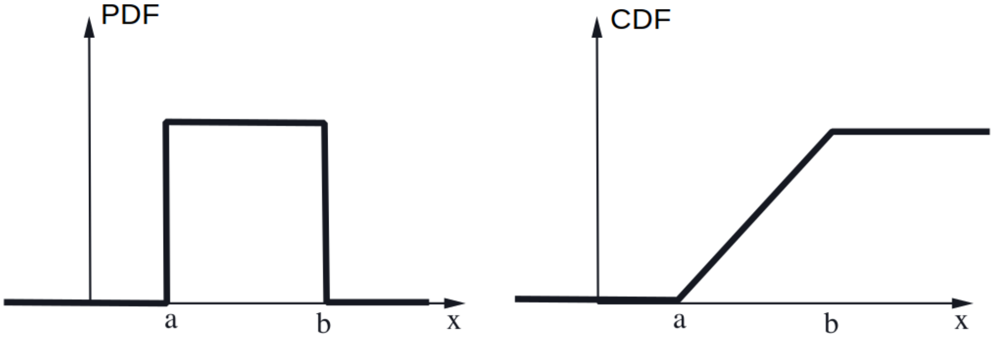
$F_X(x) = \int_{-\infty}^{x} f_X(x)dx$

For example, consider the **Uniform Distribution**, where the outcome is equally likely to occur in a fixed interval between a and b.

The PDF of the Uniform Distribution is given by,

$$\boxed{P(X=x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}}$$

The PDF and it's corresponding CDF are illustrated.



In general, there will be smoother transition in the CDF of a continuous random variable.

For the uniform distribution,

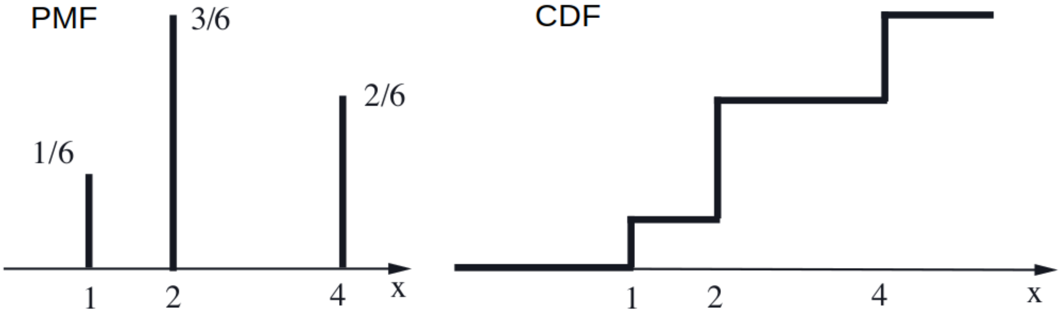
- mean, $\mu = (a+b)/2$
- variance, $\sigma^2 = (b-a)^2/12$

For discrete random variables given by PMF $p_X(x)$, the CDF will be $F_X(x) = \sum_{-\infty}^{x} p_X(x)dx$

Considering the example of rolling a dice and a discrete random variable X is defined as (where x is the number appearing on the dice),

$$X = \begin{cases} 1 & \text{if } x = 1 \\ 2 & \text{if } x = 2, 3, 4 \\ 4 & \text{if } x = 5, 6 \end{cases}$$

The PMF and CDF are illustrated.



Note that the CDF of discrete random variables will consist of sudden jumps and constant values between the jumps.

Properties of CDF

- $\lim_{x \to -\infty} F_X(x) = 0$
- $\lim_{x \to \infty} F_X(x) = 1$
- $F_X(x_1) \leq F_X(x_2) \implies x_1 < x_2$
- $P(x_1 \leq X \leq x_2) = F_X(x_2) - F_X(x_1)$

2.2.3 Joint PDF

Just like in the discrete case, if X and Y are two continuous random variables, then their joint PDF is given by, $f_{X,Y}(x, y)$. The probability of the join distribution taking on a value in the specified ranges of x and y is given by,

$$P(x, y) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x, y) dx dy$$

Probability distribution function of one random variable can be obtained from the joint distribution function by setting one of the variables to ∞. The PDF thus obtained from the joint distribution function is called Marginal distribution function.

Conditional probability in joint PDF, i.e the PDF of X taking a value in the specified range given that Y has already taken a value in it's specified range

is given by, $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$.
For this to be valid, Y should have occurred meaning $f_Y(y)$ must be a finite non-zero quantity.

The 2 events X and Y are independent if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

2.2.4 Bayes' Rule for random variables

Consider a system where the input is a random variable X, represented with the distribution $f_X(x)$ and the output is a random variable Y, represented with the distribution $f_Y(y)$.
The system properties is described by the distribution $f_{Y|X}(y|x)$ since it tells about output Y, given input X.
Bayes' rule is used to find the inferences about X given Y i.e the distribution $f_{X|Y}(x,y)$.

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)f_X(x)}{f_Y(y)}; \quad f_Y(y) = \int_x f_X(x)f_{Y|X}(y|x)dx$$

Similarly for the discrete case given by corresponding PMFs,

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)p_X(x)}{p_Y(y)}; \quad p_Y(y) = \sum_x p_X(x)p_{Y|X}(y|x)$$

Note that there might be cases where X is continuous and Y is discrete or X is discrete and Y is continuous. In these cases, the formulae need to be combined or modified accordingly (mainly just swapping between integrals and summations).

X is discrete, Y is continuous

$$p_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)p_X(x)}{p_Y(y)}; \quad f_Y(y) = \sum_x p_X(x)p_{Y|X}(y|x)$$

Example, a digital input signal is sent through a system and noise gets added to the signal.

X is continuous, Y is discrete

$$f_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)p_X(x)}{p_Y(y)}; \quad p_Y(y) = \int_x p_X(x)p_{Y|X}(y|x)dx$$

Example, a continuous signal such as the intensity of a beam of light is measured using photon count (which is discrete).

2.3 Derived Distribution Functions

If X is a discrete random variable, and Y = g(X), then to find PMF of Y, just map the favourable events in g(X) back to X.

$$p_Y(y) = P(g(X) = y) = \sum_{x:g(x)=y} p_X(x)$$

If X is a continuous random variable, and Y = g(X), then PDF of Y can be obtained by following a 2 step procedure.

- Obtain CDF of Y my reverse mapping from g(X).
  $F_Y(y) = P(Y \leq y) = P(Y \leq g(x))$
- Differentiate the CDF to obtain PDF.
  $f_y(y) = d(F_Y(y))/dy$

If X and Y are random variables such that $Y = aX + b$ where a and b are constants, then the PDF of Y is given by,

$$f_Y(y) = \frac{1}{|a|} f_X(\frac{y-b}{a})$$

This result can directly be used to solve several types of problems.

If Y = g(X) is a strictly monotonic function, meaning each value in X maps to a unique value in Y, then the PDFs of X and Y are related as,

$$\boxed{f_X(x) = f_Y(y)|\frac{dg(x)}{dx}|, \quad where\ y = g(x)}$$

2.3.1 Convolution

If X and Y are two independent discrete random variables, and if W is a random variable such that W = X + Y, then the PMF of W is given by, $p_W(w) = P(X + Y = w) = P(X = x)P(y = w - x)$

$$\implies p_W(w) = \sum_x p_X(x)p_Y(w - x)$$

This operation is called 'Convolution Sum'.

Similarly in the continuous case, the PMF of W = X + Y is,

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w-x) dx$$

This operation is called 'Convolution Integral'.

The convolution operation in general obeys Commutative, Associative and Distributive laws.

2.3.2 Covariance

Covariance is a measure of the joint variability of two random variables.
If X and Y are 2 random variables, then the covariance is given by,

$$cov(X,Y) = E[[X - E(X)][Y - E(Y)]] = E(XY) - E(X)E(Y)$$

If the greater values of one variable mainly correspond with the greater values of the other, the covariance is positive. In the opposite case, when the greater values of one variable mainly correspond to the lesser values of the other, the covariance is negative.
The sign of the covariance therefore shows the tendency in the linear relationship between the variables.

Some general results of covariance-

- $cov(X,X) = var(X)$
- if X and Y are independent events, then $cov(X,Y) = 0$
- if X or Y have zero mean, then $cov(X,Y) = E(XY)$

A utility of covariance is to find the variance of 2 dependent random variables X and Y (or more).
$var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$

The magnitude of the covariance is not easy to interpret because it is not normalized and hence depends on the magnitudes of the variables.

The normalized version of the covariance, the **Correlation Coefficient** shows by its magnitude the strength of the linear relation.
The Correlation Coefficient is represented as ρ and is given by,

$$\rho = E[(\frac{X - E(X)}{\sigma_X})(\frac{Y - E(Y)}{\sigma_Y})] = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Note that $-1 \leq \rho \leq 1$

Implications of correlation coefficient-

- $\rho = 0 \implies$ X and Y are independent events.
- $|\rho| = 1 \implies$ maximum dependence i.e $\rho = 1$ means X and Y are same, and $\rho = -1$ means X and Y share inverse relationship.

2.3.3 Conditional Expectation and Variance

Consider two random variables X and Y. The conditional expectation of X given Y takes a particular value y is given by,

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx$$

Since Y itself is a random variable, meaning the value of y is random, the expectation of X given Y is not a number, but a random variable in itself. This is because the expected value of X given Y takes some value depends on that value, so it can't be constant.

$$\implies E[X|Y] = f(Y)$$

Law of Iterated Expectations
Since $E[X|Y]$ is a random variable, it will have it's own expectation. Since $E[X|Y]$ is a function of Y, it's expectation is calculated as,

$$E[E[X|Y]] = \int_{-\infty}^{\infty} E[X|Y = y] f_Y(y) dy = E[X]$$

Therefore, the expectation of the conditional expectation of X given Y is the expectation of X itself.

Law of Total Variance
Similarly, $var(X|Y)$ is also a random variable.

- $var(X) = E(X^2) - [E(X)]^2$
  $\implies var(X|Y) = E(X^2|Y) - [E(X|Y)]^2$
- $E[var(X|Y)] = E[E(X^2|Y)] - E[[E(X|Y)]^2]$
  $\implies E[var(X|Y)] = E(X^2) - E[[E(X|Y)]^2]$
- $var(E[X|Y]) = E(E(X|Y)^2) - [E[E(X|Y)]]^2$
  $\implies var(E[X|Y]) = E(E(X|Y)^2) - [E[X]]^2$
- $\therefore var(X) = E[var(X|Y)] + var(E[X|Y])$

Sum of random number of independent random variables
N is a random number which gives the count and X_i represents the N identical random variables where each of them are independent of the rest.
Let Y be the random variable which is the sum of all X_i.
$Y = X_1 + X_2 + ... + X_N$

To find expectation of Y,
$E[Y|N = n] = E[X_1 + X_2 + ...X_N|N = n]$
$\implies E[Y|N = n] = E[X_1 + X_2 + ...X_n] = nE[X_i]$
$\therefore E[Y|N] = NE[X_i]$
Using Law of Iterated Expectations,
$E[Y] = E[E[Y|N]] = E[N[X_i]] = E[N]E[X_i]$

Similarly to find variance of Y,
$var(Y) = E[N]var(X) + (E[X])^2 var(N)$

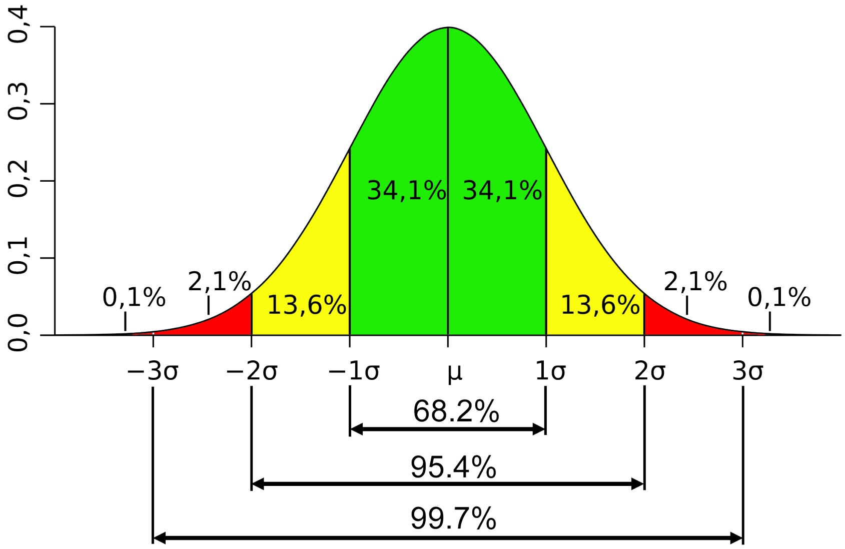
2.4 Gaussian (Normal) Distribution

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. The PDF of a General Normal Distribution is given by,

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

where μ is the mean and σ^2 is the variance.
Hence, σ is the standard deviation.

This distribution is also generally depicted as $N(\mu, \sigma^2)$



The graph illustrates an important feature of the normal distribution.

- 68.2% of the area under the curve is about the mean and between $-\sigma$ to σ.
- 95.4% of the area under the curve is about the mean and between -2σ to 2σ.
- 99.7% of the area under the curve (almost full) is about the mean and between -3σ to 3σ.

2.4.1 Standard Normal Distribution

The standard normal distribution is a specific normal distribution with zero mean and unit variance.
$\implies N(0,1)$

The PDF of a Standard Normal Distribution is given by,

$$\boxed{f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}}$$

This function is easier to deal with.

Any Gaussian distribution can effectively be converted to the standard normal distribution by performing the following substitution $z = (x - \mu)/\sigma$.

After conversion, the obtained PDF is written as $f_Z(z)$.
Also note that $P(\bar{Z}) = 1 - P(Z)$.

Some useful results of Standard Normal Distribution-

- $P(Z > 0) = P(Z < 0) = 0.5$
- $P(-1 < Z < 1) = 0.68; \quad \implies P(|Z| > 1) = 0.32$
- $P(-2 < Z < 2) = 0.95; \quad \implies P(|Z| > 2) = 0.05$
- $P(-3 < Z < 3) = 0.997; \quad \implies P(|Z| > 3) = 0.003$

3 Random Process / Stochastic Process

A random process (also called stochastic process) is a family of random variables that change with time.

3.1 Bernoulli Process

The Bernoulli process is a sequence of independent and identically distributed (iid) random variables, where each random variable takes either the value one or zero, one with probability p and zero with probability 1-p.
This process can be linked to repeatedly flipping a coin, where the probability of obtaining a head is p and its value is one, while the value of a tail is zero.
In other words, a Bernoulli process is a sequence of iid Bernoulli random variables, where each coin flip is an example of a Bernoulli trial.

At each i^{th} trial,

- $P(success) = P(X_i = 1) = p$
- $P(failure) = P(X_i = 0) = 1 - p$

The distribution associated with one such Bernoulli trial is called **Bernoulli Distribution**. The PMF of the Bernoulli Distribution is given by,
$P(X_i = x) = p^x(1-p)^{1-x}$ for x = 0 or 1.

Mean of Bernoulli Distribution, $\mu = p$
Variance of Bernoulli Distribution, $\sigma^2 = p(1-p)$

Splitting of Bernoulli Process
In a sequence of Bernoulli trials, every time a success occurs (with probability p), it is sent to one of two different sequences with probability q.
The original sequence is a Bernoulli process with parameter p and the 2 split sequences are also Bernoulli processes with parameters pq and p(1-q).
This is a result of independence of the trials.

Combining of Bernoulli Process
2 different sequences of Bernoulli trails with probabilities of success being p and q are merged into a sequence which consists of success when either of

the 2 sequences or both get a success.
The merged sequence will be a Bernoulli process with parameter
1 - (1 - p)(1 - q) = p + q - pq.
This is also a result of independence of the trials.

The Bernoulli trials have no memory. Meaning, the trials that occur after a particular set of trials have nothing to do with the previous trials.

The Bernoulli process provides a base for several discrete probability distributions.

3.1.1 Binomial Distribution

If the Bernoulli trial is conducted n times, then the random variable obtained is called Binomial Random Variable and the obtained distribution is called Binomial Distribution.

This distribution can be used to find the probability of a particular number of successes S, in the n Bernoulli trials.

Let p be the probability of success of 1 trial and probability of failure be 1 - p. If r is the number of successes required, then

$$\boxed{P(S=r) = {}^{n}C_{r}\ p^{r}\ (1-p)^{n-r};\ \ where\ r\ =\ 0,\ 1....,\ n}$$

Expectation of $P(S=r)$, $\mu = np$
Variance of $P(S=r)$, $\sigma^2 = np(1-p)$

3.1.2 Geometric Distribution

If in a Bernoulli process, if a random variable defines the number of trials till the first success, it is called Geometric Random Variable and the obtained distribution if called Geometric Distribution.

This distribution is used to find the probability that T is a particular number of trials before first success occurs.

Let p be the probability of success of 1 trial and probability of failure be 1 -

p. If t is the number of trials before first success, then

$$P(T = t) = (1-p)^{t-1}p; \quad where\ t \ = \ 1,\ 2....$$

Expectation of $P(T = t)$, $\mu = 1/p$
Variance of $P(T = t)$, $\sigma^2 = (1-p)/p^2$

3.1.3 Pascal Distribution

Consider a sequence of Bernoulli trials. Let the first success occur at $T = t_1$. The instance of occurrence of second success at $T = t_2$ will be independent of the first, and so on.
The Pascal Distribution is obtained by a random variable defined by Y_k which is the occurrence of k^{th} success at a particular instance (trial).
$Y_k = T_1 + T_2 + ...T_k$
$\implies P(Y_k = t)$ is the probability that $k-1$ successes occurring within $t-1$ trials and k^{th} success occurs at trial t.

$$P(Y_k = t) = {}^{t-1}C_{k-1}\ p^k\ (1-p)^{t-k}; \quad t \geq k$$

Expectation of $P(Y_k = t)$, $\mu = k/p$
Variance of $P(Y_k = t)$, $\sigma^2 = k(1-p)/p^2$

3.2 Poisson Process

The Poisson Process is the continuous version of the Bernoulli Process. Hence, the Poisson Process can be interpreted as the arrival of an event at some point in a continuous space.
Time Homogeneity is assumed while defining the Poisson Process i.e for any duration τ in the continuous space, the probability of arrival is the given by the same distribution.
Also, the number of arrivals in disjoint time intervals are independent.

Hence, for a very small interval δ where λ is the arrival rate, the Poisson Process is defined as,

$$P(k) = \begin{cases} \lambda & \text{if } k = 1 \\ 1-\lambda & \text{if } k = 0 \\ 0 & \text{if } k > 1 \end{cases}$$

Splitting of Poisson Process
Take a Poisson Process with arrival rate λ. Every time an arrival occurs, it is sent to one of two different sequences with probability q.
The original sequence is a Poisson process with parameter λ and the 2 split sequences are also Bernoulli processes with parameters λq and $\lambda(1-q)$.

Merging of Poisson Process
2 different Poisson Processes with arrival rates λ_1 and λ_2. Since they are independent, they can be merged into a Poisson Process with arrival rate $\lambda_1 + \lambda_2$. The probability that the arrival in the merged process is because of the first process is $\lambda_1/[\lambda_1 + \lambda_2]$ and similarly it being caused by the second process is $\lambda_2/[\lambda_1 + \lambda_2]$.

Like the Bernoulli process, the Poisson process also has no memory.
And it is used to derive several continuous distributions.

3.2.1 Poisson Distribution

In a Poisson Process, the probability of x arrivals is given by the Poisson Distribution.
This distribution is used when the number of trials are very large when compared to the probability of success.

$$\boxed{P(S = x) = \frac{\lambda^x \ e^{-\lambda}}{x!}}$$

Mean of Poisson Distribution, $\mu = \lambda$
Variance of Poisson Distribution, $\sigma^2 = \lambda$
Note that $\lambda = np$ where n is the number of trials and p is the probability of success; $n >> p$

3.2.2 Erlang Distribution

In a Poisson Process, the probability of k^{th} arrival occurring at some time is given by,

$$P(Y_k = t) = \frac{\lambda^k \; t^{k-1} \; e^{-\lambda t}}{(k-1)!}; \quad t \geq 0$$

Mean of Erlang Distribution, $\mu = k/\lambda$
Variance of Erlang Distribution, $\sigma^2 = k/\lambda^2$

3.2.3 Exponential Distribution

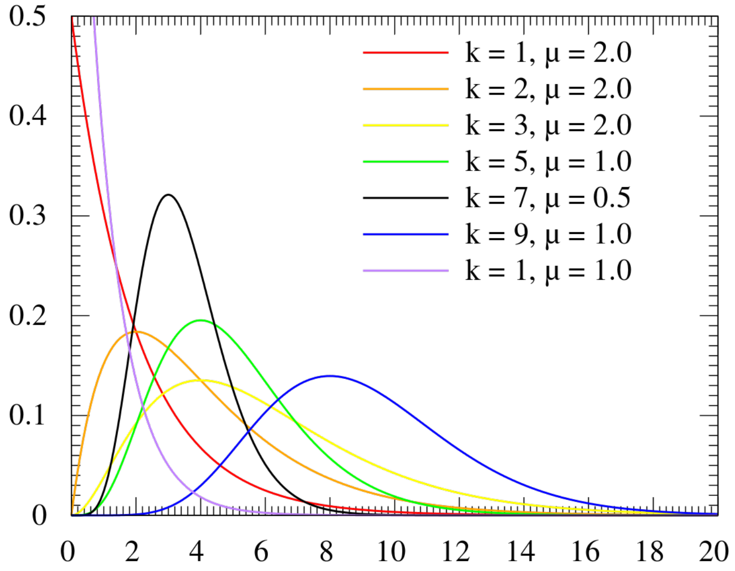
In the Erlang Distribution, if k = 1 i.e the occurrence of first arrival at some time is given by,

$$P(Y_1 = t) = \lambda \; e^{-\lambda t}; \quad t \geq 0$$

Hence, the Exponential distribution is a special case of the Erlang distribution.

Mean of Exponential Distribution, $\mu = 1/\lambda$
Variance of Exponential Distribution, $\sigma^2 = 1/\lambda^2$

Exponential and Erlang distributions for various cases are illustrated.



3.2.4 Random incidence for Poisson Process

The arrival of the required event (success) in a Poisson process is not the same as randomly choosing a particular instant. For example, if the Poisson process is a collection of infinite number durations of length 5 and 10, which are equally likely to occur.
Then the probability of choosing some duration is 0.5 and the expected value of the duration is 7.5.
However, if a random point is chosen, the expectation of the duration that it falls in is not 7.5 because since 10 is longer than 5, it is more likely that the chosen point lies in the longer duration. So here, the expected value of the duration is $10 \times 2/3 + 5 \times 1/3 = 8.333$.

3.3 Markov Chain

The Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.
A Markov chain consists of several states and transitional probabilities for each. As described, the next state depends only on the current state and does not depend the the path taken to get there.

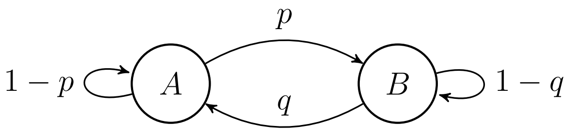
Finite Chain Markov Process
X_0 is the initial state, which is given or random.
After n transitions, the process reaches a state X_n.
There is a finite set of possible states. Let i and j be any two arbitrary states.
p_{ij} is the transition probability of moving from state i to state j and it is the conditional probability that given i, it moves to j.
$\implies p_{ij} = P(X_{n+1} = j | X_n = i)$

Model specification steps:

- identify the possible states
- identify the possible transitions
- identify the transition probabilities

The possible states are the minimum number of states required to determine the next state.

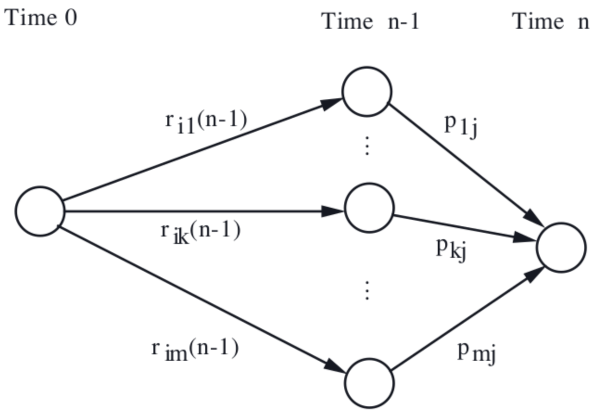
Basic example for a Markov chain is illustrated.
Consider a model with 2 states A and B. Transition from A to B and B to A occurs with probability p and q, which the probabilities that A remains at A and B remains at B are 1-p and 1-q respectively.



This sort of representation of a Markov chain is called 'State Diagram'.

n-step Transitional Probabilities

Given initial state i, after n steps, to find the probability that the chain is at state j; $r_{ij}(n) = P(X_n = j | X_0 = i)$



The required probability is the sum of all possible ways in which the process can go from i to j in n steps.
In the above diagram, note that the transition from time 0 to time n-1 can/will have various different paths and that itself will have an overall probability similar to the one of interest.
The required sum is given by Key recursion:

$$r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1)\ p_{kj}$$

This sum is given by total probability theorem after conditioning on the states just before the final (required state) i.e Time n-1. Note that the same approach can be taken but instead of taking the penultimate states for conditioning, any states in between can be considered (such as Time 1 or whatever).

3.3.1 Convergence

Every state in the Markov chain has a steady state transitional probability. This is the probability of transition in the long run.
For example, consider the Markov chain illustrated before with states A and B, with p = 0.5 and q = 0.2.

	n = 0	n = 1	n = 2	n → ∞
$r_{AA}(n)$	1	0.5	0.35	2/7
$r_{AB}(n)$	0	0.5	0.65	5/7
$r_{BA}(n)$	0	0.2	0.35	2/7
$r_{BB}(n)$	1	0.8	0.65	5/7

Convergence may or may not depend on initial state. Sometimes a case will occur where depending on the initial state, some parts of the chain are totally inaccessible (parts that would be accessible for some other initial states). In such cases, the initial condition has an effect on the convergence probabilities. Otherwise, in general convergence probability does not care about the initial state.

Recurrent and Transient States
Recurrent states are those states in a Markov chain that have return paths. Meaning, if the state is entered it can be exited and if the state is exited, it can be entered.
Transient states are those states in a Markov chain that do not have return paths. Meaning, once the state is exited, it can not occur again.

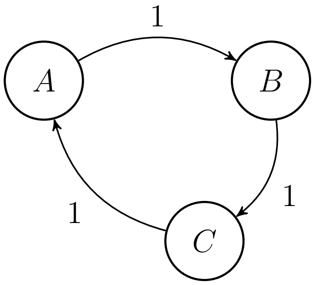
Recurrent class is a collection of recurrent states that "communicate" to each other and to no other state.

It can be noted that in the long run (steady state), the Markov process will be running around the recurrent states because the transient states would

have been exited from and they will not occur again.
Conclusion: Steady state probabilities of transient states are zero, steady state probabilities of recurrent states are non-zero.

Periodic States
The states in a recurrent class are periodic if they can be grouped into $d > 1$ groups so that all transitions from one group lead to the next group.

Consider the following state diagram example;



State A is certain to go to state B, B is certain to go to state C and C is certain to go to A. This is a simple periodic process. Here, steady state probability depends on the number of transition. Meaning, if the initial state is A, then note that

$$r_{AB}(n) = \begin{cases} 1 & \text{if } n = 1, 4, 7, ... \\ 0 & \text{otherwise} \end{cases}$$

$$r_{BC}(n) = \begin{cases} 1 & \text{if } n = 2, 5, 8, ... \\ 0 & \text{otherwise} \end{cases}$$

$$r_{CA}(n) = \begin{cases} 1 & \text{if } n = 3, 6, 9, ... \\ 0 & \text{otherwise} \end{cases}$$

In general, if a self loop exists in the state diagram, then the Markov chain is not periodic.

Based on the above concepts, it can be concluded that the transitional probabilities converge to some finite values and are independent of the initial state if -

- all the recurrent states are in a single recurrent class.
- the recurrent class does not consist of periodic states.

Since $r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1)\ p_{kj}$, to get π_j take limit as $n \to \infty$.
Hence, it becomes

$$\pi_j = \sum_{k=1}^{m} \pi_k\ p_{kj}; \quad \forall\ j$$

And sum of all steady state probabilities of a particular state is equal to 1.

$$\sum_j \pi_j = 1$$

These two equations are collectively known as **Balance Equations** and are used to solve for the steady state probabilities of a Markov chain.

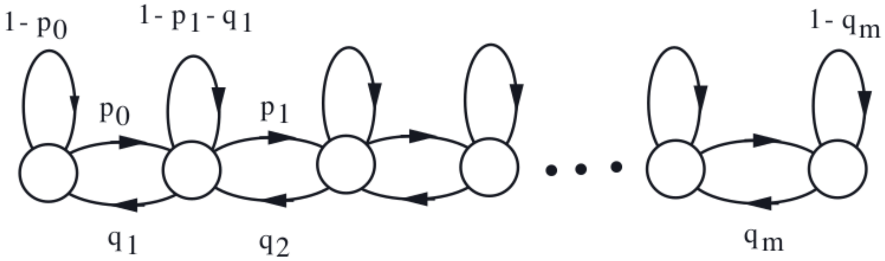
The steady state probability of a particular state can also be interpreted as the frequency of visiting that state in the long run.
Meaning, the steady state probability tells how often the Markov chain will be visiting that state.

Consider the first example again. The Balance equations will give-
$\pi_1 = 0.5\pi_1 + 0.2\pi_2; \quad \pi_2 = 0.5\pi_1 + 0.8\pi_2 \quad \Longrightarrow \ 0.5\pi_1 = 0.2\pi_2$
$\pi_1 + \pi_2 = 1$
$\therefore \pi_1 = 2/7$ and $\pi_2 = 5/7$

Note that this result is same was the results obtained before in the generic method.

3.3.2 Birth-Death Process

The birth-death process is a simple 'm+1' state Markov chain where any arbitrary state i can either remain in i or move to $i+1$ or move to $i-1$ (except for the first and last states i.e where $i = 0$, it can go to $i = 1$ or remain and where $i = m$, it can go to $i = m-1$ or remain).



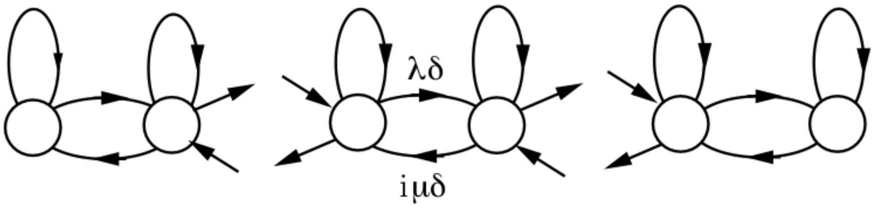
To find steady state probabilities in such a Markov chain, it is best to take some arbitrary state i and analyse it.
The state i goes to $i+1$ with probability p_i and the state $i+1$ goes back to i with probability q_{i+1}.
In the long run, the number of times transition occurs from i to $i+1$ must be almost equal to the number of transitions from $i+1$ to i.
$\implies \pi_i p_i = \pi_{i+1} q_{i+1}$
Using the above equation, all π_i can be found.

If $p_0 = p_1 = p_2 = ... + p_{m-1} = p$ and $q_1 = q_2 = q_3 = ... = q_m = q$, let $\rho = p/q$.
$\therefore \pi_i = \pi_0 \rho^i; \quad i = 0, 1, ...m$

If $p = q$, it means the states are equally likely to move forward or backward.
In this case, $\rho = 1$.
$\therefore pi_i = pi_o = 1/m + 1 \quad i = 0, 1, ...m$

If $p << q$ and $m \approx \infty$, $\pi_0 = 1 - \rho$
And steady state expectation is given by, $E[X_n] = \rho/(1-\rho)$

Phone Company Problem
This is a birth-death process, but more of a continuous time scenario divided into very small discrete intervals.
The arrival of a calls is modelled using the Poisson process, with rate of arrival λ.
The mean lifetime of a call is given by μ, which is an exponentially distributed random variable.
B lines are available. The problem is to find the optimal number of lines B such that the probability that a new customer who places a call getting busy signal (i.e all B lines being occupied) is low.



Balance equations will give, $\lambda\pi_{i-1} = \mu\pi_i$

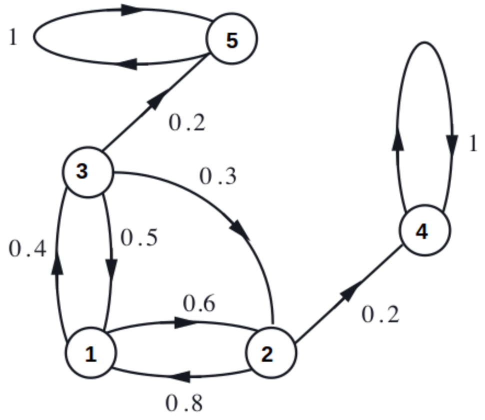
$$\implies \pi_i = \pi_0 \frac{\lambda^i}{\mu^i\ i!}; \quad \pi_0 = 1/\sum_{i=0}^{B} \frac{\lambda^i}{\mu^i\ i!}$$

3.3.3 Absorption

In a Markov chain, the transient states eventually die out. Absorption simply means the process of transient states being exited to move to recurrent states.

Absorption Probabilities
Given the Markov chain is in some transient state, the probability that it moves to a certain recurrent class is called Absorption probability.
Note that if there is only 1 recurrent class, the absorption probability is 1.
When are are multiple recurrent classes, the absorption probabilities for each of the classes have to be calculated.

Consider the following Markov chain.



Absorption probabilities are the probabilities that the Markov chain settles at state 4 and state 5 in the long run, given different initial conditions i (since the absorption probabilities depend on the initial states).

For example, consider finding the probability that the Markov chain will settle at state 4, for all possible initial states.
It can immediately be noted that $a_5 = 0$ and $a_4 = 1$.
$a_2 = 0.2 + 0.8a_1$
$a_1 = 0.4a_3 + 0.6a_2$
$a_3 = 0.3a_2 + 0.5a_1$
Solution to this system of equations will give the values of absorption probabilities of reaching state 4, given any initial state.

Similar process can be used to find the absorption probabilities for state 5.

The general set of equations would be-

$$a_i = \sum_j p_{ij} a_j \quad for\ all\ other\ i$$

In case there are multiple recurrent states in each recurrent class, the process does not change. Consider the entire recurrent class as one state and proceed with the same steps.

Expected time to Absorption
The average number of transitions required for the Markov chain to exit from it's transient states and reach one of the recurrent classes is it's expected absorption time.

For the same given Markov chain, the expected time to absorption to state 4 or 5 given initial conditions has to be found.
Direct observations, $\mu_4 = 0$ and $\mu_5 = 0$.
$\mu_2 = 1 + 0.8\mu_1$
$\mu_3 = 1 + 0.3\mu_2 + 0.5\mu_1$
$\mu_1 = 1 + 0.4\mu_2 + 0.6\mu_2$
Solution to this system of equations will give the values of expected absorption times to reach state 4, given any initial state.
1 is added because to get from that initial state to any other state, 1 transition is necessary.

Note that here, the recurrent states or classes are lumped together because expected absorption time doesn't care about which recurrent class is reached.

The general set of equations would be-

$$\mu_i = 1 + \sum_j p_{ij}\mu_j \ \ for\ all\ other\ i$$

Now consider a class of only recurrent states i.e a recurrent class.

Mean first passage time is the average number of steps required to first visit a particular state, given some initial state.

Mean first passage time from i to s:
$t_i = E[min(n \geq 0\ such\ that\ X_n = s)|X_0 = i]$

This can be solved by removing all the transitions going out of the required state because these transitions don't matter as they happen only after reaching the state (which is irrelevant for calculating the number of times needed to visit it first time).
Once that is done, the problem is essentially same as finding the mean absorption time.

Mean recurrence time is the average number of steps required to visit a particular state for the second time, given the chain was at that state initially.

Mean recurrence time of s:
$t_s = E[min(n \geq 1\ such\ that\ X_n = s)|X_0 = s]$

3.3.4 Markov Matrix