Probability and Random Variables

Sai Krishna

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: Discrete Sample spaces can be of infinite size as in, a countably infinite sample space.

1.1.1 Countably Infinite and Uncountably Infinite Sets

A set is countably infinite if its elements can be put in one-to-one correspondence with the set of natural numbers. In other words, one can count off all elements in the set in such a way that, even though the counting will take forever, you will get to any particular element in a finite amount of time.
Ex: Set of Integers

A set is uncountable if it contains so many elements that they cannot be put in one-to-one correspondence with the set of natural numbers. In other words, there is no way that one can count off all elements in the set in such a way that, even though the counting will take forever, you will get to any particular element in a finite amount of time.
Ex: The set of Real numbers

Consider the set of real numbers, given a number **a** that belongs to the set, there is no way of determining which number comes after **a**.
Assume a proposition that **b** comes after **a**. Which means there exists no number that belongs to the set of Real numbers that satisfies the following inequality

$$\mathrm{a} \leq \mathrm{x} \leq \mathrm{b}$$

But for every **b** we can define **x** as

$$\mathrm{x} = (\mathrm{a} + \mathrm{b})/2$$

Which ensures that the previous inequality is always satisfied.
Hence there is no way of determining the number that comes after **a**. Hence given any number it is not possible to count off all elements in the set until we reach that number in a finite time.

1.2 Probabilistic Models

They are mathematical description of an uncertain situation. The elements of a probabilistic model are:

- **Sample space**
  Set of all possible outcomes of an experiment
- **Probability law**
  Assigns a non negative number P(E) to every possible outcome of the experiment.
  P(E) encodes our knowledge/belief about the likelihood of the occurence of the event.

1.3 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 $E_1, E_2, ..., E_n$ i.e

  $$P(E_i \cap E_j) = \begin{cases} 0 & i \neq j \\ P(E_i) & i = j \end{cases}$$

  then,

  $$\bigcup_{i=1}^{n} P(E_i) = \sum_{i=1}^{n} P(E_i)$$

1.4 Discrete Models

1.4.1 Discrete Probability Law

If sample space contains a finite number of elements, the probability law is specified by the probabilites of **atomic events**(contain only one element).

$$\text{P(} \{E_1, E_2, ..., E_n\} \text{) = P(}E_1\text{) + P(}E_2\text{) + .. P(}E_n\text{)}$$

1.4.2 Discrete uniform law

The probability of an event in a discrete sample space is defined as,

$$P(E) = \frac{n(E)}{n(\Omega)}$$

where,
$n(E)$ is number of outcomes under event E
$n(\Omega)$ is number of outcomes in the sample space.

Ex 1: The probability of getting a number greater than 4 in a roll of a die.

$$\begin{aligned} n(\Omega) &= 6 \quad [1, 2, 3, 4, 5, 6] \\ n(E) &= 3 \quad [4, 5, 6] \\ P(E) &= \frac{3}{6} = 0.5 \end{aligned}$$

The above is an example for countable finite sample space.

Ex 2: 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 ∞.
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.

1.5 Continuous Models

- Probability of atomic events is 0
- In the interval (0,1) the probability of a sub interval (a,b) is b-a.
- P$(A \cup B \cup C)$ = P(A) + P$(A^c \cap B)$ + P$(A^c \cap B^c \cap C)$

1.5.1 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.

Ex 3: The probability of the dart landing exactly on the bulls-eye

The radius of the dart board is considered as 1
Area of the bulls eye point is 0
Area of the dart board is π.
The required probability is $\frac{0}{1} = 0$

Ex 4:The probability of the dart landing within $\frac{1}{10}^{th}$ the radius of the dart board.

Area of the required region is $\pi(0.1)^2$
Area of the dart board is π

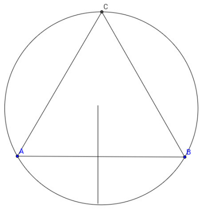
The required probability is $\frac{\pi(0.1)^2}{\pi} = 0.01$

1.6 Importance of Probability Model

BERTRAND's PARADOX

Consider a circle, and an equilateral triangle inscribed in the circle. What is the probability that the length of a randomly chosen chord of the circle is greater than the side of the triangle.

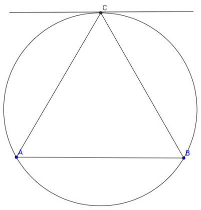
Model 1:



The radius of the circle intersects the triangle at the midpoint.
Any chord drawn perpendicular to the radius, will be longer than the side of the triangle with a probability of 0.5.
Because the side of the triangle passes exactly through the mid point of the radius.

$\therefore$ The probability is 0.5.

Model 2:



Consider the tangent drawn at one of the vertices of the equilateral triangle.
The side of the triangle is at an angle of 60^o to the tangent.
For a chord to be longer than the side of the triangle, the angle it makes with the tangent should lie in the range(60^o,120^o). Else it will be shorter than the side of the triangle.

$\therefore$ The probability is $\frac{120-60}{180} = 0.33$

Both the answers are correct according to the model defined.

Definition of the probability model is very important and influences the answer we get.

1.7 Limitations of the Probability Axiom

Consider an unit square.
Probability of one particular point on the unit square is 0.(Since the probability of atomic events in continuous case is 0).

P(Ω) = 1
In this case Ω is the unit square.

$$P(\Omega) = P(\bigcup_{x=0}^{1}\bigcup_{y=0}^{1}\{x,y\}) = \sum_{x=0}^{1}\sum_{y=0}^{1}P(\{x,y\}) = \sum_{x=0}^{1}\sum_{y=0}^{1}0 = 0$$

We just proved 1 = 0 !!

The above cannot be true,
this is due to the **limitation in the Additivity axiom**.
Additivity axiom can be used in the case when the sample space is countable.
But the number of points within a unit square is uncountably infinite.
Hence the additivity axiom cannot be used.

1.8 Impossible Event

An outcome for a given experiment which is not part of the sample space is called an Impossible event.
Let the impossible event be denoted by **I**.
Since **I** is not a part of the sample space, P(**I**) = 0.

$$\text{P}(\mathbf{I}) = 0$$
$$\text{P(E)} = 0 \nRightarrow \text{Impossible Event}$$

Ex: Probability of atomic events in the continuous case is 0. It does not mean that event is impossible.

2 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.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B) \neq 0$$

2.1 Axioms of Conditional Probability

- $P(A|B) \geq 0$
- $P(B|B) = 1$
  Since B is the new sample space.
- If events $A_1, A_2, .., A_n$ do not share any common outcomes

$$\bigcup_{i=1}^{n} P(A_i|B) = \sum_{i=1}^{n} P(A_i|B)$$

2.2 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.

2.3 Total Probability Theorem

Partition:
A collection of sets is said to be a partition of a set $\mathbb{S}$, if the sets in the collection are disjoint and their union is equal to $\mathbb{S}$.

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,

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) + ...$$

2.4 Bayes' Theorem

Provides a way to update our belief based on the arrival of new relevant pieces of information.

$$\text{Updated Belief} = \frac{\text{Likelihood x Prior Belief}}{Evidence}$$

Combining Bayes' Theorem with Total probability theorem we get

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_j P(A_j)P(B|A_j)}$$

2.5 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.

Disjoint sets ⇒ Dependent events

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.

- If P(A)P(B) = P(A ∩ B) does not always mean events are independent.

2.5.1 Conditional Independence

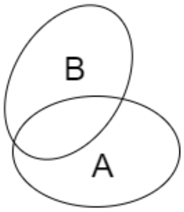
Two events A and B are conditionally independent after C has already occurred if

$$\text{P(A}\cap\text{ B|C) = P(A|C)P(B|C)}$$

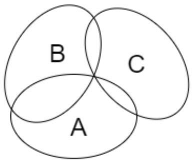
Note that independence of A and B in general does not imply conditional independence (and vice versa).

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

Ex:



Assume A,B are independent events. Given another event C has occured.



Note that A,B are disjoint given C has occured.
P$(A|C \; \cap \; B|C) = 0$
Hence they are dependent.

2.5.2 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

Ex: Consider an experiment of tossing a fair coin two times
A: First toss results in Heads → { HH, HT}
B: Second toss results in Heads → { HH, TH}
C: Both toss leads to same result → { HH, TT}

P$(A \cap B)$ = P$(A \cap C)$ = P$(B \cap C)$ = 0.25
A,B,C are pairwise independent.

P$(C|(A \cap B)) = 1 \neq$ P(C)
A,B,C are not mutually independent.

3 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.

Mapping from Sample Space to the number line is done by a Random Variable

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

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

Discrete Random Variables

Can take values from a discrete set of values.
The sample space for a discrete random variable can be discrete, continuous or even a mixture of discrete and continuous points.
Ex: Consider the dart board example, Let the number of concentric circles on it be equal to 10. We can define a discrete random variable for the above continuous sample space as

$$X = \begin{cases} 1 & \text{if dart hits ring numbers} \leq 5 \\ -1 & \text{otherwise} \end{cases}$$

Continuous Random Variables

Can take values from a continuous set of values.
A continuous random variable can only result from a continuous sample space.

3.0.1 Conditions for a Function to be a Random Variable

- Every element in the sample space must be mapped to one and only one value of the Random Variable.
- $\underset{x \to \infty}{lim} P_X(x) = 0$
- $\underset{x \to -\infty}{lim} P_X(x) = 0$

  The above two conditions does not prevent X from taking very large values (tending to ∞), it only requires that the probability of the set of those x be zero
- P(Xleqx) should be equal to the sum of the probabilities of all elementary events corresponding to { X $\leq$ x}

3.1 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,

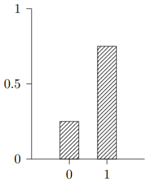
$$p_X(x) = P(X = x)$$

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.

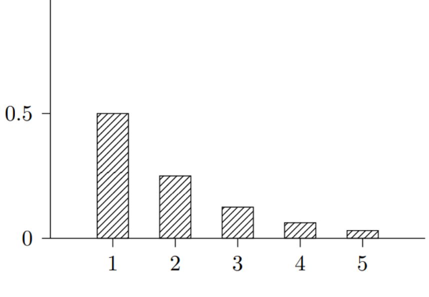
Consider another example where a fair coin is tossed twice and a random variable X is defined as,

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

.
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-

- $P_X(x) \geq 0 \implies$ each value must have non negative probability.
- $\sum\limits_x P_X(x) = 1 \implies$ sum of probabilities of all values must be equal to 1.

Cumulative Distribution Function
The probability P(X $\leq$ x) is a function of x, and is denoted as $F_X(x)$ and is called the Cumulative Distribution Function.

$$F_X(x) = P(X \leq x)$$

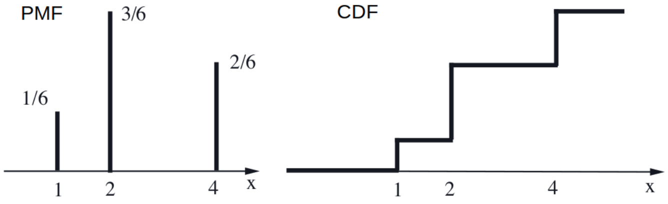
Properties:

- $\underset{x \to -\infty}{lim} F_X(x) = 0$
- $\underset{x \to \infty}{lim} F_X(x) = 1$
- $0 \leq F_X(x) \leq 1$
- $F_X(x_1) \leq F_X(x_2) \iff x_1 < x_2$
- $P(x_1 \leq X \leq x_2) = F_X(x_2) - F_X(x_1)$
- $F_X(x^+) = F_X(x)$
  CDF is a function continuous from the right

Ex: Consider a random variable X

$$P_X(x) = \begin{cases} 1/6, & x = 1 \\ 3/6, & x = 2 \\ 0, & x = 3 \\ 2/6, & x = 4 \end{cases}$$

The PMF and CDF for the above random variable is as follows



$F_X(x) = 1$ if $x > 4$

3.2 Continuous Random Variables

Continuous Random Variables are represented using Probability Density Functions.

Cumulative Distribution Function
The definition of the CDF for a continuous random variable is the same as defined for a discrete random variable.

$$F_X(x) = P(X \leq x)$$

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.

$$f_X(x) = \frac{F_X(x)}{dx}$$

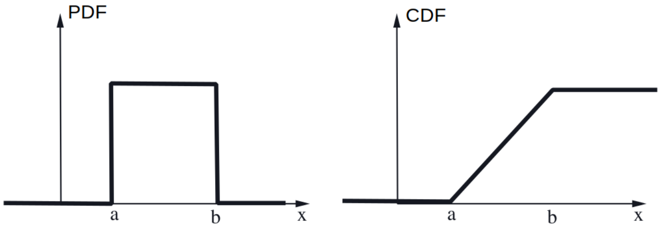
If the distribution function has sharp corners or discontinuities,then these are represented using step discontinuities or impulse functions in the plot of the density function.

Properties

- $f_X(x) \geq 0 \quad \forall x$
- $\int\limits_{-\infty}^{\infty} f_X(x).dx = 1$
- $F_X(x) = \int\limits_{-\infty}^{x} f_X(u).du$
- $P(a \leq X \leq b) = \int_a^b f_X(x).dx$

Ex: Consider a random variable which takes any value between a and b with equal probability.

The PDF and CDF of this random variable are as follows



$f_X(x) = \dfrac{1}{b-a}$ if $x \in (a, b)$
$F_X(x) = 1$ if $x > b$

3.3 Operations on Random Variable

3.3.1 Expectation

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

$$\mathbb{E}(X) = \sum_x xP_X(x)$$
$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf_X(x)dx$$

Physically it represents the centre of mass of an object, given that mass is distributed according to the given PMF/PDF.

If X is 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,

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

If $f_X(x)$ is the PDF of X and $f_Y(y)$ is the PDF of Y,
then the expectation of Y is given by,

$$\mathbb{E}(Y) = \int_y yf_Y(y).dy = \int_x g(x)f_X(x)$$

NOTE: $\mathbb{E}(g(X))$ = g($\mathbb{E}(X)$) $\implies$ g(.) is a linear function.

Let α and $\beta \in \mathbb{R}$

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

3.3.2 Moments of a Random Variable

The moments of a random variable (or of its distribution) are expected values of powers or related functions of the random variable.
n^{th} moment of a Random Variable is defined as

$$\mathbb{E}(X^n) = \sum_x x^n P_X(x)$$
$$\mathbb{E}(X^n) = \int_x x^n f_X(x).dx$$

Mean value of a random variable is the first moment of the random variable.

3.3.3 Central Moments of a Random Variable

n^{th} central moment of a Random Variable is defined as

$$\mathbb{E}\big(X - \mathbb{E}(X)\big)^n$$

The first central moment is always equal to 0.

3.3.4 Variance

It is a measure of the spread of the random numbers from their average value. It is the second central moment of a random variable.

$$var(X) = \mathbb{E}\big(X - \mathbb{E}(X)\big)^2 = \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2$$

Let α and $\beta \in \mathbb{R}$

- $var(\alpha) = 0$
- $var(\alpha X) = \alpha^2 var(X)$
- $var(\alpha X + \beta) = \alpha^2\ var(X)$

Note: $var(X)$ is always a positive quantity.

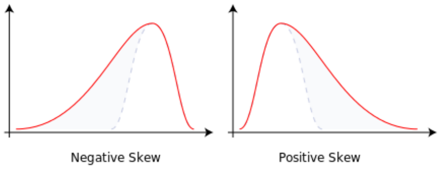
Standard deviation of X is defined as the positive square root of the variance of X and is denoted by σ_X. $\implies \sigma_X = \sqrt{var(X)}$

3.3.5 Skew

The third central moment of a random variable is a measure of the asymmetry of the probability distribution function about its mean.

- **Negative Skew**
  The left tail is longer; the mass of the distribution is concentrated on the right of the mean.

- **Positive Skew**
  The right tail is longer; the mass of the distribution is concentrated on the right of the mean.



3.3.6 Markov's Inequality

Gives an upper bound for the probability that a non negative function of a random variable is greater than or equal to some positive constant.

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a} \qquad a > 0$$

Proof:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf_X(x).dx = \int_{-\infty}^{a} xf_X(x).dx + \int_{a}^{\infty} xf_X(x).dx \geq \int_{a}^{\infty} xf_X(x).dx$$

$$\int_{a}^{\infty} xf_X(x).dx \geq \int_{a}^{\infty} af_X(x).dx$$

$$\implies \mathbb{E}(X) \geq \int_{a}^{\infty} af_X(x).dx \; = \; aP(X \geq a)$$

3.3.7 Chebyshev's Inequality

Guarantees that for a wide class of probability distributions, no more than a certain fraction of values can be more than a certain distance from the mean.

$$P(\; |X - \mathbb{E}(X)| \; \geq \; k) \; \leq \; \frac{var(X)}{k^2}$$

Let $\mathbb{E}(X) = \mu$
Proof:

$$P((X-\mu)^2 > k^2) \leq \frac{\mathbb{E}((X-\mu)^2)}{k^2} \text{ From Markov Inequality}$$

$$P(|X-\mu| > k) \leq \frac{var(X)}{k^2}$$

The above two inequalities hold for the discrete case as well. The proof for which is similar to the above ones.(Replace PDF with PMF)

3.3.8 Characteristic function of Random Variables

The characteristic function of a random variable is given by

$$\Phi_X(\omega) = \mathbb{E}\left[e^{j\omega X}\right]$$

Expressing the Expectation operator interms of the density function we get

$$\Phi_X(\omega) = \int_{-\infty}^{\infty} f_X(x) e^{j\omega x}.dx$$

$\Phi_X(\omega)$ is seen to be the Fourier Transform of $f_X(x)$ with the sign of ω reversed.

3.3.9 Moment Generating Function

The moment generating function of a random variable is given by

$$M_X(t) = \mathbb{E}\left[e^{tX}\right]$$

To generate the nth moment of a random variable, differentiate the MGF n times and evaluate it at t=0.

$$\frac{d^n}{dt^n} M_X(t)\Bigg|_{t=0} = \mathbb{E}(X^n)$$

3.3.10 Transformation of a Random Variable

Monotonic Transformation of a Continuous Random Variable

Consider a monotonic transformation **T**. **T** is both continuous and differentiable for all values of x for which $f_X(x) \neq 0$.

Consider a particular value y_0 corresponding to x_0.

$$y_0 = T(x_0) \quad \implies x_0 = T^{-1}(y_0)$$

CASE 1: T is a monotonically increasing function.
Hence probability of the event $\{Y \leq y_0\}$ is equal to the probability of the event $\{X \leq x_0\}$.

$$\begin{aligned} F_Y(y_0) &= F_X(x_0) \\ F_Y(y_0) &= F_X(T^{-1}(y_0)) \\ \int_{-\infty}^{x_0} f_X(x).dx &= \int_{-\infty}^{T^{-1}(y_0)} f_X(x).dx \end{aligned}$$

Leibnitz Rule of differentiation.

$$\frac{d}{dy_0}\Big(\int_{-\infty}^{y_0} f_Y(y).dy\Big) = f_Y(y_0).\frac{dy_0}{dy_0} - f_Y(-\infty)\frac{d}{dy_0}(-\infty) + \int_{-\infty}^{y_0} \frac{\partial f_Y(y)}{\partial y_0}.dy$$

Differentiate on both sides with respect to y_0 using Leibnitz rule of differentiation

$$f_Y(y_0) = f_X(T^{-1}(y_0))\frac{dT^{-1}(y_0)}{dy_0}$$

The result applies for any value of x_0 and y_0

$$f_Y(y) = f_X(T^{-1}(y))\frac{dT^{-1}(y)}{dy}$$
$$f_Y(y) = f_X(x)\frac{dx}{dy}$$

CASE 2: T is a monotonically decreasing function.
Hence probability of the event $\{Y \leq y_0\}$ is equal to the probability of the event $\{X \geq x_0\}$.

$$F_Y(y_0) = 1 - F_X(x_0)$$
$$F_Y(y_0) = 1 - F_X(T^{-1}(y_0))$$
$$\int_{-\infty}^{x_0} f_X(x).dx = 1 - \int_{-\infty}^{T^{-1}(y_0)} f_X(x).dx$$

Following the same steps as CASE 1, we will end up with the same result with a negative sign, but since the slope of a monotonically decreasing function is negative. The overall sign is positive. Hence we get the same result for both the cases.

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

Nonmonotonic Transformation of a Continuous Random Variable

The derivation is similar to the previous case, but we have to consider multiple regions of x which correspond to $\{Y \leq y_0\}$

$$f_Y(y) = \sum_n f_X(x_n)\left|\frac{dx}{dT(x)\Big|_{x=x_n}}\right|$$

x_n includes all the regions of x, which correspond to $\{Y \leq y_0\}$

4 Multiple Random Variables

4.1 Joint Distribution function

The probability of the joint event $\{X \leq x, Y \leq y\}$ which is a function of both the numbers x and y, is called the joint probability distribution function and denoted by $F_{XY}(x,y)$

$$F_{XY}(x,y) = P(X \leq x, Y \leq y)$$

Properties

- $F_{XY}(-\infty, -\infty) = F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0$
- $F_{XY}(\infty, \infty) = 1$
- $F_{XY}(x,y)$ is a non decreasing function of both x and y
- $F_{XY}(x_2,y_2) + F_{XY}(x_1,y_1) - F_{XY}(x_1,y_2) - F_{XY}(x_2,y_1) =$
  $P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2)$
- $F_{XY}(\infty, y) = F_Y(y) \quad F_{XY}(x, \infty) = F_X(x)$

4.1.1 Marginal Distribution function

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.

4.2 Joint Density function

It is the second derivative of the joint distribution function.

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$$

Properties

- $\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$
- $F_{XY}(x,y) = \int\limits_{-\infty}^{y} \int\limits_{-\infty}^{x} f_{X,Y}(x,y)dxdy$
- $F_X(x) = \int\limits_{-\infty}^{x} \int\limits_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy$

  $F_Y(y) = \int\limits_{-\infty}^{y} \int\limits_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy$
- $P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int\limits_{y_1}^{y_2} \int\limits_{x_1}^{x_2} f_{X,Y}(x,y)dxdy$

4.2.1 Marginal Density function

Probability Density function of one random variable can be obtained from the joint density function

$$f_X(x) = \frac{dF_X(x)}{dx} = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

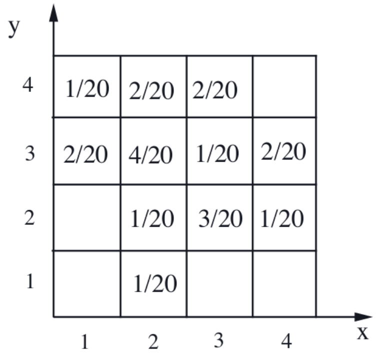
4.3 Joint PMF

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,

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

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.

4.3.1 Marginal PMF

Individual PMF can also be found from joint PMF.

$$P_X(x) = \sum_y P_{X,Y}(x,y)$$
$$P_Y(y) = \sum_x P_{X,Y}(x,y)$$

A few examples from the given joint PMF,
$P_{X,Y}(1,3) = 2/20$
$P_X(3) = \sum\limits_y P_{XY}(3,y) = 0 + 3/20 + 1/20 + 2/20 = 6/20$
$P_Y(1) = 1/20.$

4.4 Conditional Distribution and Density function

The conditional distribution function of a random variable X, given some event B has occured is given by

$$F_X(x|B) = P(X \leq x|B) = \frac{P(X \leq x \cap B)}{P(B)}$$

Defining the event B as

$$B = \{y - \Delta y \ < \ Y \ < \ y + \Delta y\}$$

$$F_X(x|y - \Delta y \ < \ Y \ < \ y + \Delta y) = \frac{\int_{y-\Delta y}^{y+\Delta y} \int_{-\infty}^{x} f_{X,Y}(u,v)dudv}{\int_{y-\Delta y}^{y+\Delta y} f_Y(v)dv}$$

The above is called **Point conditioning** as we let Δy approach 0.

$$F_X(x|Y = y) = \frac{\int_{-\infty}^{x} f_{X,Y}(u,y)du}{f_Y(y)}$$

Differentiating on both sides(Using Leibnitz rule) we get the conditional density function

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

If we define B as, (assuming P(B) $\neq$ 0)

$$B = \{y_a \ < \ Y \ \leq \ y_b\}$$

$$F_X(x|y_a \ < \ Y \ \leq \ y_b) = \frac{F_{X,Y}(x,y_b) - F_{X,Y}(x,y_a)}{F_Y(y_b) - F_Y(y_a)}$$

$$f_X(x|y_a \ < \ Y \ \leq \ y_b) = \frac{\int_{y_a}^{y_b} f_{X,Y}(x,y)dy}{\int_{y_a}^{y_b} \int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy}$$

The above is called **Interval Conditioning**

4.4.1 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.
From the previous example considered for joint PMF

- $P_{X|Y}(2|3) = P(X = 2|Y = 3) = P(X = 2 \cap Y = 3|Y = 3) = \dfrac{4/20}{9/20} = \frac{4}{9}$
- 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

4.5 Statistical Independence of Random Variables

The same condition of independence used with respect to events, can also be used to define the statistical independence of two random variables.

Two random variables are statistically independent if

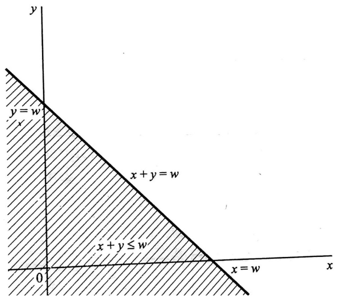
$$\begin{aligned} P(X \le x, Y \le y) &= P(X \le x)P(Y \le y) \\ F_{X,Y}(x,y) &= F_X(x)F_Y(y) \\ f_{X,Y}(x,y) &= f_X(x)f_Y(y) \end{aligned}$$

- In case of independence conditional distribution/density functions are same as marginal distribution/density functions.

4.6 Sum of Independent Random Variables

Let X,Y be independent random variables, and W be another random variable defined as

$$\begin{aligned} W &= X + Y \\ F_W(w) = P(W \le w) &= P(X + Y \ \le w) \end{aligned}$$



The shaded region gives the required probability.

$$\begin{aligned} F_W(w) &= \int_{-\infty}^{\infty}\int_{-\infty}^{w-y} f_{X,Y}(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{w-y} f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} f_Y(y)F_X(w-y)dy \end{aligned}$$

Differentiate on both sides using Leibnitz rule wrt w

$$f_W(w) = \int_{-\infty}^{\infty} f_Y(y)f_X(w-y)dy$$

The above expression is the convolution integral.

Density function of the sum of two independent random variables is the convolution of their individual density functions.

The above analysis can be extended to sum of n independent random variables.

$$\begin{aligned} Y &= X_1 + X_2 + .. + X_n \\ f_Y(y) &= f_{X_1}(x_1) * f_{X_2}(x_2) * ... * f_{X_n}(x_n) \end{aligned}$$

The same analysis holds true for the case of discrete random variables as well.

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

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

4.7 Operations on Multiple Random Variables

4.7.1 Expectation and Variance

When more than a single random variable is involved, expectation must be taken with respect to all the random variables involved.

$$\mathbb{E}(g(X_1, X_2, .., X_n)) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} g(x_1, x_2, .., x_n) f_{X_1,X_2,..,X_n}(x_1, x_2, .., x_n)dx_1dx_2..dx_n$$

In general, expectation is a linear operation, meaning it satisfies

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

- If X and Y are independent, then $\mathbb{E}$(XY) = $\mathbb{E}$(X)$\mathbb{E}$(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

$$\mathbb{E}[\mathrm{g(X)h(Y)}] = \mathbb{E}[\mathrm{g(X)}]\mathbb{E}[\mathrm{h(Y)}]$$

- $var(aX + bY) = a^2var(X) + b^2var(Y) + 2abCov(X, Y)$
- In case of independence of X and Y, the variances satisfy

$$\mathrm{var(X + Y) = var(X) + var(Y)}$$

- If Z = X - kY, then $var(Z) = var(X) + k^2var(Y)$

4.7.2 Joint Moments about Origin

They are denoted as m_{nk}

$$m_{nk} = \mathbb{E}(X^nY^k) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^ny^kf_{X,Y}(x, y)dxdy$$

- m_{n0} are the moments of X.
  m_{0k} are the moments of Y.
- The sum n + k is called the order of the moments.
  m_{10}, m_{01} are the first order moments
  m_{20}, m_{02}, m_{11} are the second order moments

4.7.3 Joint Central Moments

They are denoted as μ_{nk}

Let $\overline{X}$,$\overline{Y}$ be the mean of X,Y

$$\begin{aligned}\mu_{nk} &= \mathbb{E}[(X - \overline{X})(Y - \overline{Y})] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - \overline{X})^n(y - \overline{Y})^kf_{X,Y}(x, y)dxdy\end{aligned}$$

4.7.4 Correlation

The second order moment m_{11} is called the correlation of X and Y.
It is denoted as R_{XY}

$$R_{XY} = m_{11} = \mathbb{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x, y)dxdy$$

- X,Y are uncorrelated or independent $\implies \quad R_{XY} = \mathbb{E}(X)\mathbb{E}(Y)$

- Independence $\Rightarrow$ Uncorrelatedness
  Uncorrelatedness $\not\Rightarrow$ Independence
- X,Y are orthogonal $\implies \quad R_{XY} = 0$

Ex: Let X be a random variable with mean 3, and variance 2.
Let Y = -6X + 22

$\mathbb{E}(Y) = \mathbb{E}(-6X + 22) = -6\mathbb{E}(X) + 22 = 4$
$\mathbb{E}(X^2) = var(X) + (\mathbb{E}(X))^2 = 11$
$R_{XY} = \mathbb{E}(XY) = \mathbb{E}(-6X^2 + 22X) = -6\mathbb{E}(X^2) + 22\mathbb{E}(X) = 0$

From the above we notice that X,Y are orthogonal though they are correlated.
NOTE: If Y = aX + b

- $|a| \neq 0 \implies$ X,Y are correlated
- $b = \dfrac{-a\mathbb{E}(X^2)}{\mathbb{E}(X)} \implies$ X,Y are orthogonal.
- If $\mathbb{E}(X) = 0$, X and Y are orthogonal $\iff$ a = 0

4.7.5 Covariance

Covariance is a measure of the joint variability of two random variables. The second order moment μ_{11} is called the covariance of X and Y.
It is denoted as C_{XY}

$$C_{XY} = \mu_{11} = \mathbb{E}[(X - \bar{X})(Y - \bar{Y})] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - \bar{X})(y - \bar{Y})f_{X,Y}(x, y)dxdy$$
$$C_{XY} = R_{XY} - \mathbb{E}(X)\mathbb{E}(Y)$$

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.

- X,Y are uncorrelated or independent $\implies \quad C_{XY} = 0$
- X,Y are orthogonal $\implies \quad C_{XY} = -\mathbb{E}(X)\mathbb{E}(Y)$
- Correlation coefficient of X and Y is $\rho_{XY} = \dfrac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}} \qquad -1 \leq \rho_{XY} \leq 1$
  - $\rho_{XY} = 0 \implies$ X,Y are independent
  - $\rho_{XY} = 1 \implies$ X,Y are same
  - $\rho_{XY} =$ -1 $\implies$ X,Y share inverse relationship
- $Cov(X, X) = var(X)$

4.7.6 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.

$$\mathbb{E}[\ \mathbb{E}(g(X)|Y)\ ] = \mathbb{E}[\ g(X)\ ]$$

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

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

NOTE: The definition of the above operations holds for the discrete case as well.

5 Standard Probability distributions

5.1 Bernoulli distribution

Discrete probability distribution of a random variable which can take only two values.

$$\boxed{P_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}} \quad 0 < p < 1$$

- Mean = p
- Variance = p(1-p)
- MGF = 1 - p + pe^t

5.2 Binomial distribution

Discrete probability distribution that models the number of successes in a sequence of n independent Bernoulli trials.

$$\boxed{P_X(k) = {}^nC_k P^k (1-p)^{n-k}}$$

- Mean = np
- Variance = np(1-p)
- MGF = $(1 - p + pe^t)^n$

5.3 Geometric distribution

Discrete probability distribution of a random variable that defines the number of independent Bernoulli trials required till the first success is achieved.
The probability of success in each trial being p.

$$\boxed{P_X(k) = (1-p)^k p}$$

The above is the definition of geometric distribution used for modeling the number of trials up to and including the first success.

- Mean = $\dfrac{1-p}{p}$
- Variance = $\dfrac{1-p}{p^2}$
- MGF = $\dfrac{p}{1-(1-p)e^t}$

Alternatively,

$$\boxed{P_X(k) = (1-p)^{k-1}p}$$

The above is the definiton of geometric distribution used for modeling the number of failures until the first success.

- Mean $= \dfrac{1}{p}$
- Variance $= \dfrac{1-p}{p^2}$
- MGF $= \dfrac{pe^t}{1-(1-p)e^t}$

5.4 Negative Binomial distribution

Discrete probability distribution that models the number of failures in a sequence of i.i.d Bernoulli trials before a specified (non random) number of success occurs.
Let $r > 0$ denote the number of success expected
p denote the success in each trial

$$\boxed{P_X(k,r) = \ {}^{k+r-1}C_k \ p^k(1-p)^r}$$

Ex: Define rolling a 6 on a die as a success, and rolling any other number as a failure, and ask how many failed rolls will occur before we see the third success (r = 3). In such a case, the probability distribution of the number of non-6s that appear will be a negative binomial distribution.

- Mean $= \dfrac{pr}{1-p}$
- Variance $= \dfrac{pr}{(1-p)^2}$
- MGF $= \big(\dfrac{1-p}{1-pe^t}\big)^r$ for t < -logp

NOTE: If r(stopping parameter) is an integer then negative binomial is used. If r is a real number then **Polya distribution** is used.

5.5 Pascal Distribution

Special case of negative binomial 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 X_k which

is the occurrence of k^{th} success at a particular instance (trial).
$X_k = T_1 + T_2 + ...T_k$
$\implies P(X_k = t)$ is the probability that $k-1$ successes occurring within $t-1$ trials and k^{th} success occurs at trial t.

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

- Mean $= \dfrac{k}{p}$
- Variance $= \dfrac{k(1-p)}{p^2}$

Substituting t as r+k we get the negative binomial distribution.

5.6 Poisson distribution

Discrete probability distribution that expresses the probability of a given number of events occuring in a fixed interval of time or space if the events occur with a known constant mean rate and independently since the time of occurence of the last event.

NOTE: The limiting of the binomial distribution with $p \to 0$ and n being a very large number is the Poisson distribution.

$$\boxed{P_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}}$$

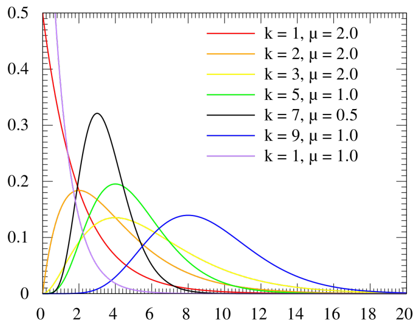
- Mean $= \lambda =$ np
- Variance $= \lambda =$ np
- MGF $= e^{[\lambda(e^t-1)]}$

5.7 Erlang distribution

A specific case of the gamma distribution. Used in queuing theory to calculate the waiting time.

$$\boxed{P_X(k,\mu) = \frac{x^{k-1}e^{-x/\mu}}{\mu^k\ k!}}$$

k is the shape parameter (positive integer).
μ is the scale parameter (positive real number)



- Mean = kμ
- Variance = kμ^2
- MGF = $\big(1 - \frac{t}{\lambda}\big)^{-k}$ for $t < \lambda$

5.8 Exponential Distribution

In the Erlang Distribution, if k = 1 we get the exponential distribution.

$$\boxed{f_X(x) = \lambda e^{-\lambda x}}$$

- Mean = $1/\lambda$
- Variance = $1/\lambda^2$
- MGF = $\dfrac{\lambda}{\lambda - t}$

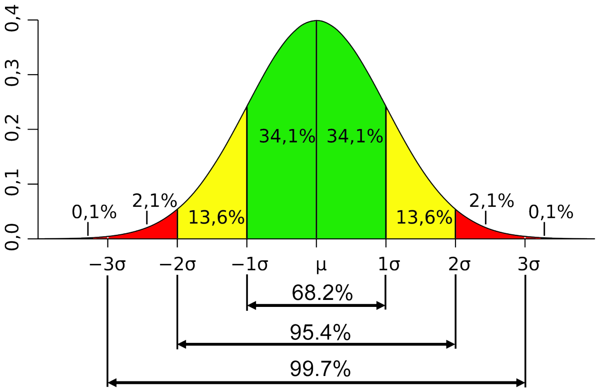
5.9 Normal distribution

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.

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

where,
μ is the mean and σ^2 is the variance of the distribution

- MGF = $\exp\big(\mu t + \dfrac{\sigma^2 t^2}{2}\big)$



5.9.1 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-

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

5.9.2 Properties of Gaussian random variable

- Completely defined only through their first and second moments
- If a pair of Gaussian random variables are uncorrelated then they are also independent
- Random variables produced by a linear transformation of normal random variables is also a normal random variable.
- Conditional density of gaussian random variables is also gaussian

5.10 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).