Probability Theory refresher

Sample Space

Sample Space: The set of all possible outcomes of an experiment is called the sample space and is denoted by Ω.
Individual elements are denoted by ω and are termed elementary outcomes.
Examples:

- (Finite) A single roll of an ordinary die. Here, $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- (Countable) Infinite number of coin tosses in order to study, say, the number of tosses before 5 consecutive heads are observed. Here, $\Omega = \{H, T\}^\infty$.
- (Uncountable) Speed of a vehicle measured with infinite precision. Here, $\Omega = \mathbb{R}$.

Event

Event: An event is any collection of possible outcomes of an experiment, that is, any subset of Ω.

In most experiments we are generally more interested in observing the occurrence of particular events rather than the elementary outcomes. For example, on rolling a die, we may be interested in observing whether the outcome was even (event $E = \{2, 4, 6\}$) or odd (event $O = \{1, 3, 5\}$).

Set Theory Notations

$$A \subset B \Leftrightarrow x \in A \Rightarrow x \in B$$

$$A = B \Leftrightarrow A \subset B \text{ and } B \subset A$$

$$A \cup B = \{x : x \in A \text{ or } x \in B\}$$

$$A \cap B = \{x : x \in A \text{ and } x \in B\}$$

$$A^c = \{x : x \notin A\}$$

Properties of Set Operations

Commutativity

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

Associativity

$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A \cap (B \cap C) = (A \cap B) \cap C$$

Distributivity

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

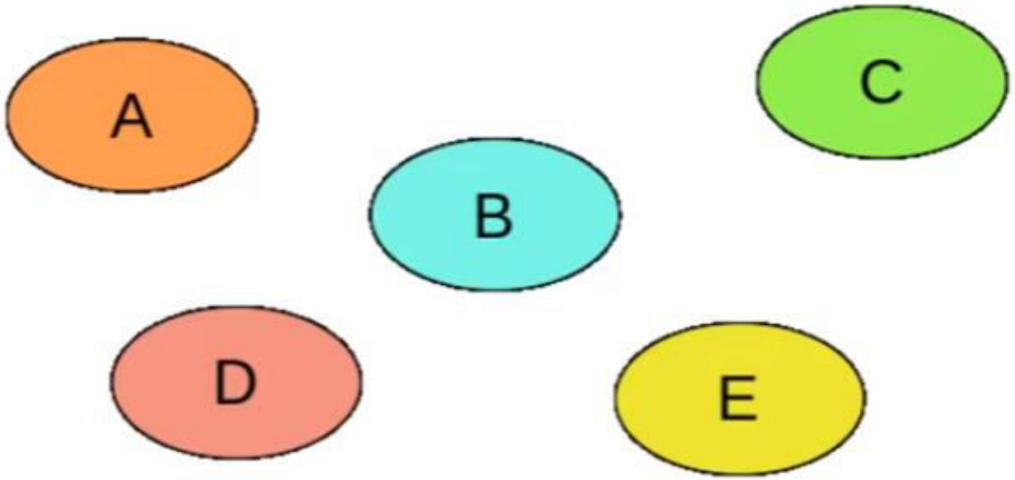
DeMorgan's Laws

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Disjoint Events

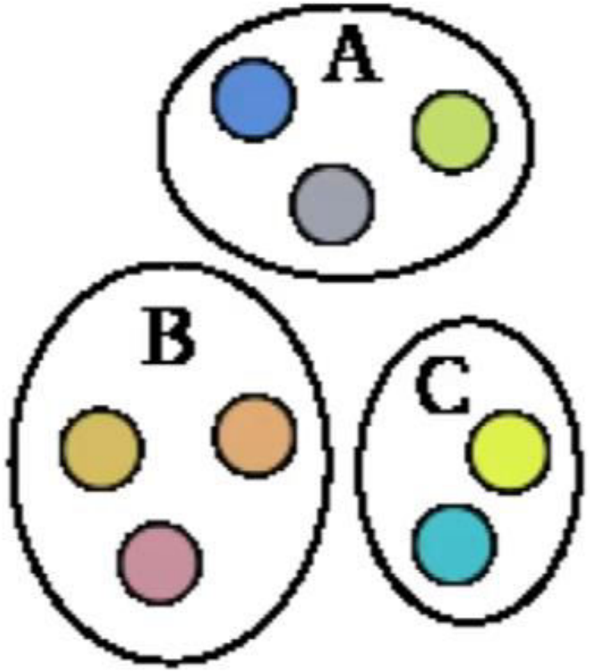
Two events A and B are disjoint (or mutually exclusive) if $A \cap B = \phi$.

A sequence of events $A_1, A_2, A_3, \ldots$ are pair-wise disjoint if $A_i \cap A_j = \phi$ for all $i \neq j$.



Partition

If $A_1, A_2, \ldots$ are pair-wise disjoint and $\cup_{i=1}^{\infty} A_i = \Omega$, then the collection $A_1, A_2, \ldots$ forms a partition of Ω.



Sigma Algebra

Given a sample space Ω, a σ-algebra is a collection $\mathcal{F}$ of subsets of Ω, with the following properties:

(a) $\Phi \in \mathcal{F}$.
(b) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
(c) If $A_i \in \mathcal{F}$ for every $i \in \mathbb{N}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

A set A that belongs to $\mathcal{F}$ is called an $\mathcal{F}$-measurable set (event).

Example: Consider $\Omega = \{1, 2, 3\}$.
$\mathcal{F}_1 = \{\phi, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$.
$\mathcal{F}_2 = \{\phi, \{1, 2, 3\}\}$.

Sample Space Size Considerations

For any Ω (countable or uncountable) 2^{Ω} is always a σ-algebra.

For example, for $\Omega = \{H, T\}$, a feasible σ-algebra is the power set, i.e., $\mathcal{F} = \{\phi, \{H\}, \{T\}, \{H, T\}\}$.

However, if Ω is uncountable, then probabilities cannot be assigned to every subset of 2^{Ω}.

Probability Measure & Probability Space

A probability measure $\mathcal{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathcal{P} : \mathcal{F} \to [0, 1]$ satisfying

(a) $\mathcal{P}(\phi) = 0, \quad \mathcal{P}(\Omega) = 1;$
(b) if $A_1, A_2, ...$ is a collection of pair-wise disjoint members of $\mathcal{F}$, then

$$\mathcal{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathcal{P}(A_i)$$

The triple $(\Omega, \mathcal{F}, \mathcal{P})$, comprising a set Ω, a σ-algebra $\mathcal{F}$ of subsets of Ω, and a probability measure $\mathcal{P}$ on $(\Omega, \mathcal{F})$, is called a **probability space**.

Example

Consider a simple experiment of rolling an ordinary die in which we want to identify whether the outcome results in a prime number or not.

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$\mathcal{F} = \{\phi, \{1, 4, 6\}, \{2, 3, 5\}, \{1, 2, 3, 4, 5, 6\}\}$$

$$\mathcal{P} : \mathcal{F} \rightarrow [0, 1]$$

- $\mathcal{P}(\phi) = 0$
- $\mathcal{P}(\{1, 4, 6\}) = 0.5$
- $\mathcal{P}(\{2, 3, 5\}) = 0.5$
- $\mathcal{P}(\Omega) = 1$

Bonferroni's Inequality

$$P(A \cap B) \geq P(A) + P(B) - 1$$

General form:

$$P(\cap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$$

Gives a lower bound on the intersection probability which is useful when this probability is hard to calculate.

Only useful if the probabilities of individual events are sufficiently large.

Boole's Inequality

$P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$, for **any** sets $A_1, A_2, ...$

Gives a useful upper bound for the probability of the union of events.

Conditional Probability

Given two events A and B, if $P(B) > 0$, then the conditional probability that A occurs given that B occurs is defined to be

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

Essentially, since event B has occurred, it becomes the new sample space.

Conditional probabilities are useful when reasoning in the sense that once we have observed some event, our beliefs or predictions of related events can be updated/improved.

Example

Q. A fair coin is tossed twice. What is the probability that both tosses result in heads given that at least one of the tosses resulted in a heads?

Sol. $\Omega = \{HH, TT, HT, TH\}$
$\mathcal{P}(HH) = \mathcal{P}(TT) = \mathcal{P}(HT) = \mathcal{P}(TH) = 1/4$

$$\begin{aligned}
&\mathcal{P}(HH|\text{at least one toss heads}) \\
&\quad = \mathcal{P}(HH|HT \cup TH \cup HH) \\
&\quad = \frac{\mathcal{P}(HH \cap (HT \cup TH \cup HH))}{\mathcal{P}(HT \cup TH \cup HH)} \\
&\quad = \frac{\mathcal{P}(HH)}{\mathcal{P}(HT \cup TH \cup HH)} \\
&\quad = \frac{1}{3}
\end{aligned}$$

Bayes' Rule

We have:

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

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

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

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

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(B|A)\mathcal{P}(A)}{\mathcal{P}(B)} \text{ (Bayes' Rule)}$$

Bayes' Rule

Let $A_1, A_2, \ldots$ be a partition of the sample space, and let B be any subset of the sample space. Then, for each i = 1, 2, ...,

$$\mathcal{P}(A_i|B) = \frac{\mathcal{P}(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} \mathcal{P}(B|A_j)\mathcal{P}(A_j)}$$

Bayes' rule is important in that it allows us to compute the conditional probability $\mathcal{P}(A|B)$ from the "inverse" conditional probability $\mathcal{P}(B|A)$.

Example

Q. To answer a multiple choice question, a student may either know the answer or may guess it. Assume that with probability p the student knows the answer to a question, and with probability q, the student guesses the right answer to a question she does not know. What is the probability that for a question the student answers correctly, she actually knew the answer to the question?

Example

Q. To answer a multiple choice question, a student may either know the answer or may guess it. Assume that with probability p the student knows the answer to a question, and with probability q, the student guesses the right answer to a question she does not know. What is the probability that for a question the student answers correctly, she actually knew the answer to the question?

Sol. Let K be the event that the student knows the question, and C be the event that the student answers the question correctly.
We have $\mathcal{P}(K) = p$, $\mathcal{P}(\neg K) = 1 - p$, $\mathcal{P}(C|K) = 1$, $\mathcal{P}(C|\neg K) = q$

$$\begin{aligned}
\mathcal{P}(K|C) &= \frac{\mathcal{P}(C|K)\mathcal{P}(K)}{\mathcal{P}(C)} \\
&= \frac{\mathcal{P}(C|K)\mathcal{P}(K)}{\mathcal{P}(K)\mathcal{P}(C|K)+\mathcal{P}(\neg K)\mathcal{P}(C|\neg K)} \\
&= \frac{p}{p+q(1-p)}
\end{aligned}$$

Independent Events

Two events, A and B, are said to be independent if

$$\mathcal{P}(A \cap B) = \mathcal{P}(A)\mathcal{P}(B)$$

More generally, a family $A_i : i \in I$ is called independent if

$$\mathcal{P}(\cap_{i \in J} A_i) = \prod_{i \in J} \mathcal{P}(A_i)$$

for all finite subsets J of I.

From the above, it should be clear that pair-wise independence does not imply independence.

Conditional Independence

Let A, B, and C be three events with $\mathcal{P}(C) > 0$. The events A and B are called conditionally independent *given* C if

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

or equivalently

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

Example: Assume that admission into the M.Tech. programme at IITM & IITB is based solely on candidate's GATE score. Then

$$\mathcal{P}(IITM|IITB \cap GATE) = \mathcal{P}(IITM|GATE)$$

Random Variable

A random variable is a function $X : \Omega \to \mathbb{R}$, i.e., it is a function from the sample space to the real numbers.

Examples:

- The sum of outcomes on rolling 3 dice.
- The number of heads observed when tossing a fair coin 3 times.

Induced Probability Function

Consider the previous example experiment of tossing a fair coin 3 times. Let X be the number of heads obtained in the three tosses. Enumerating the elementary outcomes, we observe the value of X as

ω	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(\omega)$	3	2	2	2	1	1	1	0

Instead of using the probability measure defined on the elementary outcomes or events, we would ideally like to measure the probability of the random variable taking on values in its range.

x	0	1	2	3
$\mathcal{P}_X(X = x)$	1/8	3/8	3/8	1/8

Induced Probability Function

Let $\Omega = \{\omega_1, \omega_2, ...\}$ be a sample space and $\mathcal{P}$ be a probability measure (function).

Let X be a random variable with range $\mathcal{X} = \{x_1, x_2, ..., x_m\}$.

We define the induced probability function $\mathcal{P}_X$ on $\mathcal{X}$ as

$$\mathcal{P}_X(X = x_i) = \mathcal{P}(\{\omega_j \in \Omega : X(\omega_j) = x_i\})$$

Cumulative Distribution Function

The cumulative distribution function or cdf of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = \mathcal{P}_X(X \leq x), \text{ for all } x$$

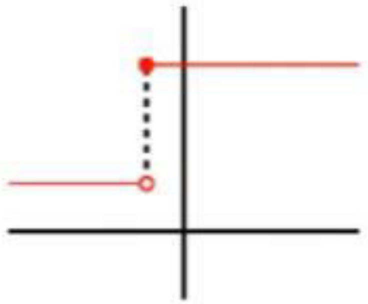
Example:

x	$(-\infty, 0]$	$(-\infty, 1]$	$(-\infty, 2]$	$(-\infty, 3]$	$(-\infty, \infty)$
$F_X(x)$	$1/8$	$1/2$	$7/8$	1	1

Properties of cdf

A function $F_X(x)$ is a cdf iff the following three conditions hold:

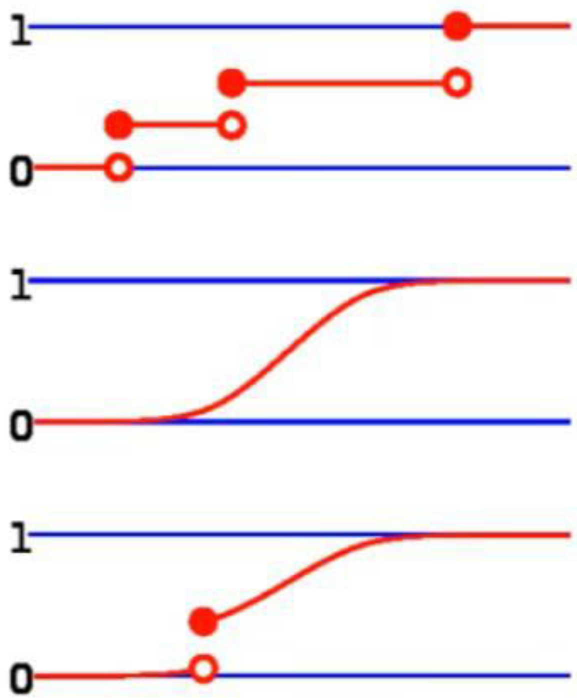
- (Monotonicity) If $x \leq y$, then $F_X(x) \leq F_X(y)$
- (Limiting values) $\lim_{x\to-\infty} F_X(x) = 0$ and $lim_{x\to\infty} F_X(x) = 1$
- (Right-continuity) For every x, we have $\lim_{y\downarrow x} F_X(y) = F_X(x)$



Continuous & Discrete Random Variables

A random variable X is continuous if $F_X(x)$ is a continuous function of x.

A random variable X is discrete if $F_X(x)$ is a step function of x.



Probability Mass Function

The probability mass function or pmf of a discrete random variable X is given by

$$f_X(x) = \mathcal{P}(X = x), \text{ for all x}$$

Example: For a geometric random variable X with parameter p,

$$f_X(x) = \begin{cases} (1-p)^{x-1}p & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Properties:

- $f_X(x) \geq 0$, for all x
- $\sum_x f_X(x) = 1$

Probability Density Function

The probability density function or pdf of a continuous random variable is the function $f_X(x)$ which satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \text{ for all x}$$

Properties:

- $f_X(x) \geq 0$, for all x
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$

Expectation

The expected value or mean of a random variable X, denoted by $E[X]$, is given by

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \text{ (continuous RV)}$$

$$E[X] = \sum_{x:\mathcal{P}(x)>0} x f_X(x) = \sum_{x:\mathcal{P}(x)>0} x \mathcal{P}(X = x) \text{ (discrete RV)}$$

Example

Q. Let the random variable X take values -2, -1, 1, 3 with probabilities $1/4$, $1/8$, $1/4$, $3/8$ respectively. What is the expectation of the random variable $Y = X^2$?

Sol. The random variable Y takes on the values 1, 4, 9 with probabilities $3/8$, $1/4$, $3/8$ respectively.
Hence,

$$E(Y) = \sum_x x\mathcal{P}(Y = x) = 1 \cdot \frac{3}{8} + 4 \cdot \frac{1}{4} + 9 \cdot \frac{3}{8} = \frac{19}{4}$$

Alternatively,

$$E(Y) = E(X^2) = \sum_x x^2\mathcal{P}(X = x) = 4 \cdot \frac{1}{4} + 1 \cdot \frac{1}{8} + 1 \cdot \frac{1}{4} + 9 \cdot \frac{3}{8} = \frac{19}{4}$$

Properties of Expectations

Let X be a random variable and let a, b, c be constants. Then, for functions $g_1(X)$ and $g_2(X)$ whose expectations exist

- $E(ag_1(X) + bg_2(X) + c) = aEg_1(X) + bEg_2(X) + c$
- If $g_1(X) \geq 0$ for all x, then $Eg_1(X) \geq 0$
- If $g_1(X) \geq g_2(X)$ for all x, then $Eg_1(X) \geq Eg_2(X)$
- If $a \leq g_1(X) \leq b$, for all x, then $a \leq Eg_1(X) \leq b$

Moments

For each integer n, the n^{th} moment of X is

$$\mu'_n = EX^n$$

The n^{th} central moment of X is

$$\mu_n = E(X - \mu)^n$$

Variance

The variance of a random variable X is its second central moment.
$VarX = E(X - \mu)^2 = E(X - EX)^2 = EX^2 - (EX)^2$

The positive square root of $VarX$ is the standard deviation of X.

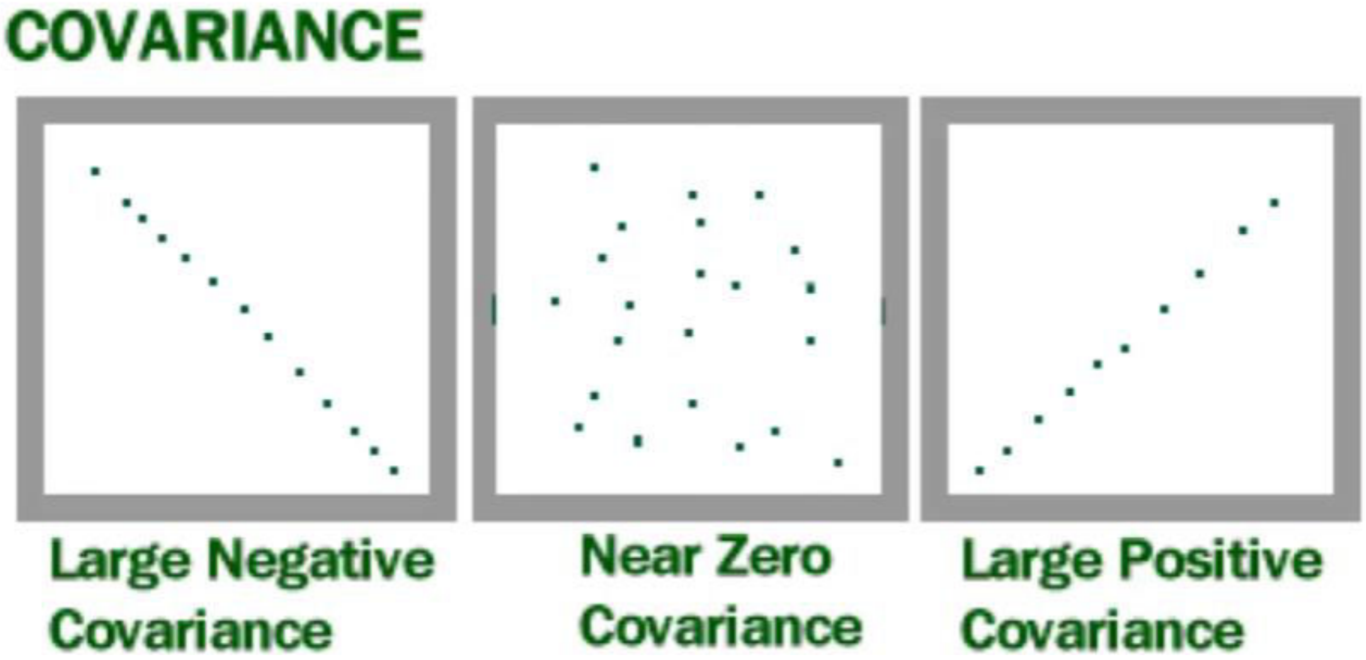
Note: $Var(aX + b) = a^2 VarX$
where a, b are constants

Covariance

The covariance of two random variables, X and Y is

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

It is a measure of how much two random variables change together.



Correlation

The correlation of two random variables, X and Y is

$$\rho(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}$$

Note:

- For correlation to be defined, individual variances must be non-zero and finite
- $\rho(X, Y)$ lies between -1 and $+1$

Probability Distributions

Consider two variables X and Y, and suppose we know the corresponding probability mass functions f_X and f_Y

Can we answer the following question:

$$\mathcal{P}(X = x \text{ and } Y = y) = ?$$

Joint Distributions

To capture the properties of two random variables X and Y, we use the joint PMF

$f_{X,Y} : \mathbb{R}^2 \to [0,1]$, defined by
$f_{X,Y}(x,y) = \mathcal{P}(X = x, Y = y)$

Marginal Distributions

Suppose we are given the joint PMF

$$f_{X,Y}(x,y) = \mathcal{P}(X = x, Y = y)$$

From this joint PMF, we can obtain the PMF's of the two random variables

$f_X = \sum_y f_{X,Y}(x,y)$ (marginal PMF of R.V. X)
$f_Y = \sum_x f_{X,Y}(x,y)$ (marginal PMF of R.V. Y)

Conditional Distributions

Like joint distributions, we can also consider conditional distributions

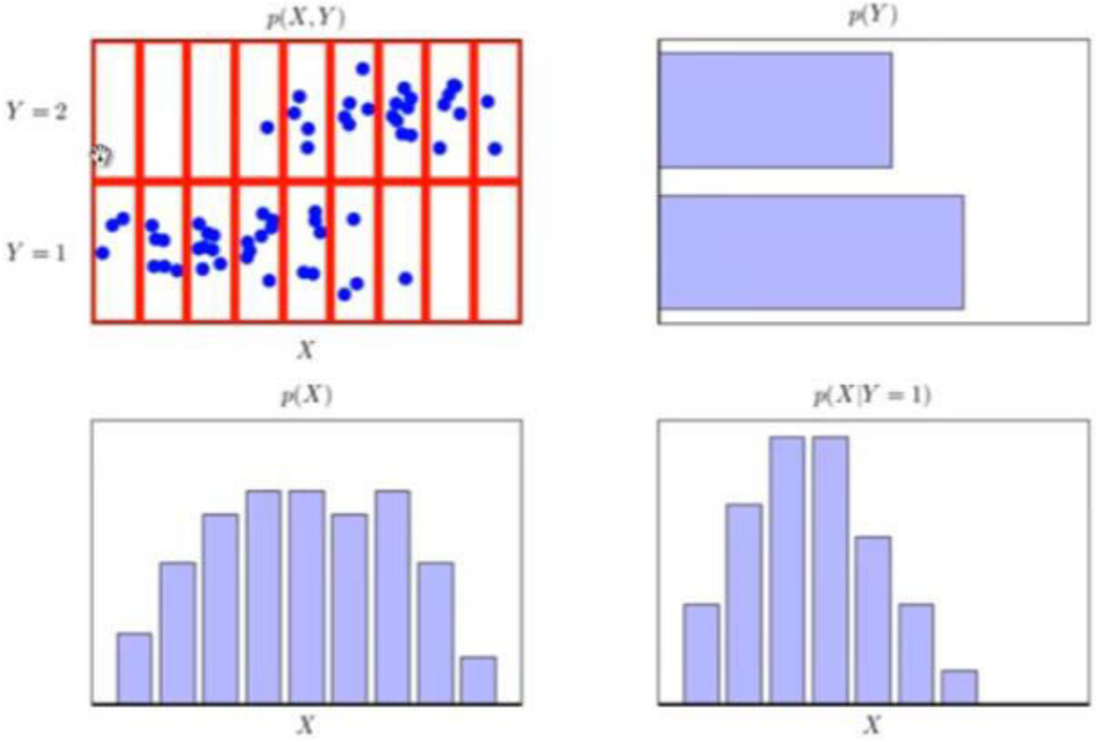
$$f_{X|Y}(x|y) = \mathcal{P}(X = x|Y = y)$$

Using conditional probability definition, we have

$$f_{X|Y}(x|y) = f_{X,Y}(x, y)/f_Y(y)$$

Note that the above conditional probability is undefined if $f_Y(y) = 0$.

Example



Bernoulli Distribution

Consider a random variable X taking one of two possible values (either 0 or 1). Let the PMF of X be given by

$$
\begin{aligned}
f_X(0) &= \mathcal{P}(X = 0) = 1 - p \qquad\qquad (0 \leq p \leq 1) \\
f_X(1) &= \mathcal{P}(X = 1) = p
\end{aligned}
$$

This describes a Bernoulli distribution

$$
\begin{aligned}
E[X] &= p \\
var(X) &= p(1 - p)
\end{aligned}
$$

Binomial Distribution

Consider the situation where we perform n independent Bernoulli trials where

- probability of success (for each trial) $= p$
- probability of failure $= 1 - p$

Let X be the number of successes in the n trials, then we have

$$\mathcal{P}(X = x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $\binom{n}{x} = \frac{n!}{(n-x)!x!}$
and $0 \leq x \leq n$

$E[X] = np$
$var(X) = np(1-p)$

Geometric Distribution

Suppose we perform a series of independent Bernoulli trials, each with a probability p of success. Let X represent the number of trials before the first success, then we have

$$\mathcal{P}(X = x|p) = (1-p)^{x-1}p \qquad x = 1, 2, 3, \ldots$$

$$E[X] = 1/p$$
$$var(X) = (1-p)/p^2$$

Uniform Distribution

A continuous random variable X is said to be uniformly distributed on an interval $[a, b]$ if its PDF is given by

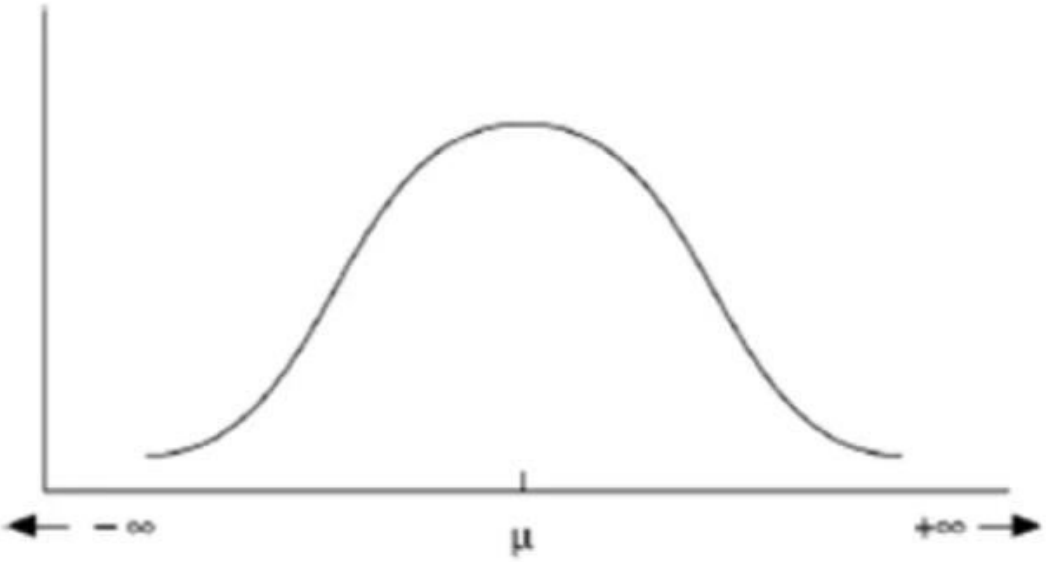
$$f_X(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$E[X] = (a+b)/2$
$var(X) = (b-a)^2/12$

Normal Distribution

A continuous random variable X is said to be normally distributed with parameters μ and σ^2 if the density of X is given by

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



Importance of Normal Distribution

Roughly, the central limit theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.

Multivariate Normal Distribution

$$\mathcal{N}(x|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} exp(-\frac{1}{2}(x-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}))$$

where

- $\boldsymbol{\mu}$ is the D-dimensional mean vector,
- $\boldsymbol{\Sigma}$ is the $D \times D$ covariance matrix, and
- $|\boldsymbol{\Sigma}|$ is the the determinant of the covariance matrix

Beta Distribution

The pdf of the beta distribution in the range $[0, 1]$, with shape parameters α, β, is given by

$$f(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where the gamma function is an extension of the factorial function.

$E[X] = \alpha/(\alpha+\beta)$
$var(X) = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$

Beta Distribution