Review of Basic Probability

P.S. Sastry
sastry@iisc.ac.in

Reference Material

- V.K. Rohatgi and A.K.Md.E. Saleh, An Introduction to probability and Statistics, Wiley, 2nd edition, 2018
- S.Ross, ‘Introduction to Probability Models’, Elsevier, 12th edition, 2019.
- P G Hoel, S Port and C Stone, Introduction to Probability Theory, 1971.
- Scott Sheffield, Probability and Random Variables, Massachusetts Institute of Technology, MIT OpenCourseWare: https://ocw.mit.edu/courses/mathematics/18-600-probability-and-random-variables-fall-2019/ License: Creative Commons BY-NC-SA.

Probability Theory

- Probability Theory – branch of mathematics that deals with modeling and analysis of random phenomena.
- Random (or Chance) Phenomena
  – individually not predictable but have a lot of regularity at a population level (e.g., tossing a coin)
- Recommender systems, opinion polls, sample surveys $\cdots$
  – useful because at a population level customer behaviour can be predicted.
- Statistics is the branch of Maths that deals with making inferences from data and Probability theory is needed for that.
- It is useful in many engineering systems.

- There are many situations where one needs to deal with random phenomena.
  - Analysis of dynamical systems subjected to noise
  - System estimation
  - Policies for decision making under uncertainty
  - Pattern Recognition, prediction from data

    ⋮
- We may use probability models for analysing algorithms. (e.g., average case complexity of algorithms)
- We may deliberately introduce randomness in an algorithm
  (e.g., ALOHA protocol, Primality testing)

  ⋮

This is only a 'sample' of possible application scenarios!

We assume all of you are familiar with the terms:

random experiment, sample space, events etc.

We use the following Notation:

- Sample space – Ω
  (outcomes of the random experiment)
  We write $\Omega = \{\omega_1, \omega_2, \cdots\}$ when it is countable
- An event is, by definition, a subset of Ω
- Set of all possible events: $\mathcal{F} \subseteq 2^\Omega$ (power set of Ω)
  We can take $\mathcal{F} = 2^\Omega$ (every subset of Ω is an event)
  (We always assume that $\mathcal{F}$ is closed under countable unions, intersections and complements. Such a $\mathcal{F}$ is called a σ-algebra).

Probability axioms

Probability (or probability measure) is a function that assigns a number in $[0, 1]$ to each event and satisfies some properties.

Formally, $P : \mathcal{F} \to \Re$ satisfying

- A1 Non-negativity: $P(A) \geq 0$, $\forall A \in \mathcal{F}$
- A2 Normalization: $P(\Omega) = 1$,
- A3 σ-additivity: If $A_1, A_2, \cdots \in \mathcal{F}$ satisfy $A_i \cap A_j = \phi, \forall i \neq j$ then

$$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i), \forall n; \quad \text{and} \quad P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Events satisfying $A_i \cap A_j = \phi, \forall i \neq j$ are said to be **mutually exclusive**

$(\Omega, \mathcal{F}, P)$ is called the **Probability Space**

Probability Space

A probability model is specified by: $(\Omega, \mathcal{F}, P)$
Ω is Sample Space; $\mathcal{F} \subseteq 2^{\Omega}$ is the set of events, and
$P : \mathcal{F} \rightarrow [0, 1]$, with

A1 $P(A) \geq 0, \forall A \in \mathcal{F}$

A2 $P(\Omega) = 1$

A3 If $A_i \cap A_j = \phi, \forall i \neq j$ then $P(\cup_i A_i) = \sum_i P(A_i)$

- Note that to specify a model we need to specify P.

Case of Countable Ω

- Let $\Omega = \{\omega_1, \omega_2, \cdots\}$ (finite or countably infinite).
- Let $q_i, i = 1, 2, \cdots$ be numbers such that $q_i \geq 0$ and $\sum_i q_i = 1$.
- We now set $P(\{\omega_i\}) = q_i, i = 1, 2, \cdots$.
- If $A = \{\omega_1, \omega_2\}$ then
  $A = \{\omega_1\} \cup \{\omega_2\}$ (mutually exclusive).
  Hence,
  $P(A) = P(\{\omega_1\}) + P(\{\omega_2\})$.
- Thus for any A: $P(A) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{i:\omega_i \in A} q_i$
- Assumptions on q_i needed to satisfy $P(A) \geq 0$ and $P(\Omega) = 1$.
- This is how we normally specify probability measure (for countable Ω).

Simple example

- If $|\Omega| = n$, we can take $q_i = \frac{1}{n}, \forall i$.
  ("All outcomes are equally likely")
- Then $P(A) = \sum_{\omega \in A} P(\{\omega\}) = \frac{|A|}{|\Omega|}$
  ("favourable divided by total number of outcomes")
- A simple example:
  - tossing three coins, "equally likely" outcomes
  - This gives:
    $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
  - If A is "getting 2 head", then
    $A = \{HHT, HTH, THH\}$ and $P(A) = \frac{3}{8}$

- For countable $\Omega = \{\omega_1, \omega_2, \cdots\}$, we can assign P using $P(\{\omega_i\}) = q_i$ where $q_i \geq 0$, $\sum_i q_i = 1$.
- If Ω is finite we can take "equally likely". $(q_i = \frac{1}{n})$ Though it is not necessary to always do this.
- Thus we know how to assign P if Ω is finite or countably infinite.
- Ω can be uncountably infinite.
- Simple idea of extending "equally likely": if $\Omega \subset \Re$ then $P(A) = \frac{|A|}{|\Omega|}$ where $|A|$ is length of A. (In general we specify such P as distributions of continuous random variables).

Example: Uncountably infinite Ω

Problem: A rod of unit length is broken at two random points. What is the probability that the three pieces so formed would make a triangle.

- Let us take left end of the rod as origin and let x, y denote the two successive points where the rod is broken.
- Then the random experiment is picking two numbers x, y with $0 < x < y < 1$.
- We can take $\Omega = \{(x, y) \ : \ 0 < x < y < 1\} \subset \Re^2$.
- For the pieces to make a triangle, sum of lengths of any two should be more than the third.

- The lengths are: $x, (y-x), (1-y)$. So we need

$$x + (y-x) > (1-y) \;\Rightarrow\; y > 0.5$$

$$x + (1-y) > (y-x) \;\Rightarrow\; y < x + 0.5;$$

$$(y-x) + 1 - y > x \;\Rightarrow\; x < 0.5$$

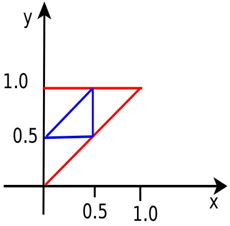
- So the event of interest is:

$$A = \{(x, y)\ :\ y > 0.5;\ x < 0.5;\ y < x{+}0.5,\ 0 < x, y < 1\}$$

- We have

$$\begin{aligned} \Omega &= \{(x, y) \ : \ 0 < x < y < 1\} \\ A &= \{(x, y) \in \Omega \ : \ y > 0.5; \ x < 0.5; \ y < x + 0.5\} \end{aligned}$$

- We can visualize it as follows



- The required probability is area of A divided by area of Ω which gives the answer as 0.25

Some Simple consequences of the axioms

(Notation: A^c is complement of A)

- $P(A^c) = 1 - P(A)$ for all events A.
- Let $A \subseteq B$. Then

$$P(A) \leq P(B), \qquad P(B - A) = P(B) - P(A)$$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  and

$$P(U_{i=1}^{n} A_i) \;=\; \sum_i P(A_i) - \sum_i \sum_{j>i} P(A_i \cap A_j)$$
$$+ \sum_i \sum_{j>i} \sum_{k>j} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} P(\cap_i A_i)$$

Known as inclusion-exclusion formula

Conditional Probability

- Let B be an event with $P(B) > 0$. We define conditional probability of any event A, conditioned on B, as

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

- The above is a notation. "$A \mid$ B" does not represent any set operation! (Maybe an abuse of notation!)
- Given a B, conditional probability is a new probability assignment to any event.
- That is, given $(\Omega, \mathcal{F}, P)$, $B \in \mathcal{F}$ with $P(B) > 0$, we define a new probability $P_B : \mathcal{F} \to [0, 1]$ by

$$P_B(A) = \frac{P(AB)}{P(B)}$$

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

- Note $P(B|B) = 1$ and $P(A|B) > 0$ only if $P(AB) > 0$.
- Now the 'new' probability of each event is determined by what it has in common with B.
- If we know the event B has occurred, then based on this knowledge we can readjust probabilities of all events and that is given by the conditional probability.
- Intuitively it is as if the sample space is now reduced to B because we are given the information that B has occurred.
- This is a useful intuition and we use this often for calculating conditional probability.

- In a conditional probability, the conditioning event can be any event (with positive probability)
- In particular, it could be intersection of events.
- We think of that as conditioning on multiple events.

$$P(A \mid B, C) = P(A \mid BC) = \frac{P(ABC)}{P(BC)}$$

- The conditional probability is defined by

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

- This gives us a useful identity

$$P(AB) = P(A \mid B)P(B)$$

- We can iterate this for multiple events

$$P(ABC) = P(A \mid BC)P(BC) = P(A \mid BC)P(B \mid C)P(C)$$

This is a very useful identity.

- Let $B_1, \cdots, B_m$ be events such that $\cup_{i=1}^{m} B_i = \Omega$ and $B_i B_j = \phi, \forall i \neq j$.
- Such a collection of events is said to be a partition of Ω. (They are also sometimes said to be mutually exclusive and collectively exhaustive).
- Given this partition, any other event can be represented as a mutually exclusive union as

$$A = AB_1 + \cdots + AB_m$$

(Notation $A = B + C$ means $A = B \cup C$ and B, C are mutually exclusive)

$$A = A \cap \Omega = A \cap (B_1 \cup \cdots \cup B_m) = (A \cap B_1) \cup \cdots \cup (A \cap B_m)$$

Hence, $A = AB_1 + \cdots + AB_m$

Total Probability rule

- Let $B_1, \cdots, B_m$ be a partition of Ω.
- Then, for any event A, we have

$$\begin{aligned} P(A) &= P(AB_1 + \cdots + AB_m) \\ &= P(AB_1) + \cdots + P(AB_m) \\ &= P(A \,|B_1)P(B_1) + \cdots + P(A \,|B_m)P(B_m) \end{aligned}$$

- The formula (where B_i form a partition)

$$P(A) = \sum_i P(A \mid B_i)P(B_i)$$

  is known as **total probability rule** or total probability law or total probability formula.
- This is a very useful in many situations. ("arguing by cases")

Example: Polya's Urn

An urn contains r red balls and b black balls. We draw a ball at random, note its color, and put back that ball along with c balls of the same color. We keep repeating this process. Let R_n (B_n) denote the event of drawing a red (black) ball at the n^{th} draw. We want to calculate the probabilities of all these events.

- It is easy to see that $P(R_1) = \frac{r}{r+b}$ and $P(B_1) = \frac{b}{r+b}$.
- For R_2 we have, using total probability rule,

$$
\begin{aligned}
P(R_2) &= P(R_2 \mid R_1)P(R_1) + P(R_2 \mid B_1)P(B_1) \\
&= \frac{r+c}{r+c+b}\,\frac{r}{r+b} + \frac{r}{r+b+c}\,\frac{b}{r+b} \\
&= \frac{r(r+c+b)}{(r+c+b)(r+b)} = \frac{r}{r+b} = P(R_1)
\end{aligned}
$$

- Similarly we can show that $P(B_2) = P(B_1)$.
- One can show by mathematical induction that $P(R_n) = P(R_1)$ and $P(B_n) = P(B_1)$ forall n. (Left as an exercise for you!)
- This does not depend on the value of c!

Bayes Rule

- Another important formula based on conditional probability is Bayes Rule:

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

- This allows one to calculate $P(A \mid B)$ if we know $P(B \mid A)$.
- Useful in many applications because one conditional probability may be more easier to obtain (or estimate) than the other.
- Often one uses total probability rule to calculate the denominator in the RHS above:

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

Example: Bayes Rule

Let D and D^c denote someone being diagnosed as having a disease or not having it. Let T_+ and T_- denote the events of a test for it being positive or negative. (Note that $T_+^c = T_-$). We want to calculate $P(D|T_+)$.

- We have, by Bayes rule,

$$P(D|T_+) = \frac{P(T_+|D)P(D)}{P(T_+|D)P(D) + P(T_+|D^c)P(D^c)}$$

- The probabilities $P(T_+|D)$ and $P(T_+|D^c)$ can be obtained through, for example, experiments.
- $P(T_+|D)$ is called the true positive rate and $P(T_+|D^c)$ is called false positive rate.
- We also need $P(D)$, the probability of a random person having the disease.

- Let us take some specific numbers
- Let: $P(D) = 0.5$, $P(T_+|D) = 0.99$, $P(T_+|D^c) = 0.05$.
$$P(D|T_+) = \frac{0.99 * 0.5}{0.99 * 0.5 + 0.05 * 0.5} = 0.95$$
That is pretty good.
- But taking $P(D) = 0.5$ is not realistic. Let us take $P(D) = 0.1$.
$$P(D|T_+) = \frac{0.99 * 0.1}{0.99 * 0.1 + 0.05 * 0.9} = 0.69$$
- Now suppose we can improve the test so that $P(T_+|D^c) = 0.01$
$$P(D|T_+) = \frac{0.99 * 0.1}{0.99 * 0.1 + 0.01 * 0.9} = 0.92$$
- These different cases are important in understanding the role of false positives rate.

- Bayes rule can be used in 'non-binary' situations also
- Let $B_1 + B_2 + B_3 = \Omega$.

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{\sum_{j=1}^{3} P(A|B_j)P(B_j)}$$

- Example: I have three coins with probability of heads being 0.1, 0.5, 0.8. I choose one at random and toss it twice and see heads both times. What is the probability it is the fair coin?
- Useful in many applications

Independent Events

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

$$P(AB) = P(A)P(B)$$

- Note that this is a definition. Two events are independent if and only if they satisfy the above.
- Suppose $P(A), P(B) > 0$. Then, if they are independent

$$P(A|B) = \frac{P(AB)}{P(B)} = P(A); \text{ similarly } P(B|A) = P(B)$$

- This gives an intuitive feel for independence.
- Independence is an important (often confusing!) concept.

Example: Independence

A class has 20 female and 30 male course (MTech) students and 6 female and 9 male research (PhD) students. Are gender and degree independent?

- Let F, M, C, R denote events of female, male, course, research students
- From the given numbers, we can easily calculate the following:

$$P(F) = \frac{26}{65} = \frac{2}{5};\ P(C) = \frac{50}{65} = \frac{10}{13};\ P(FC) = \frac{20}{65} = \frac{4}{13}$$

- Hence we can verify

$$P(F)P(C) = \frac{2}{5}\,\frac{10}{13} = \frac{4}{13} = P(FC)$$

and conclude that F and C are independent.
Similarly we can show for others.

- In this example, if we keep all other numbers same but change the number of male research students to, say, 12 then the independence no longer holds.
  $(\frac{26}{68} \, \frac{50}{68} \neq \frac{20}{68})$
- One needs to be careful about independence!
- We always have an underlying probability space $(\Omega, \mathcal{F}, P)$
- Once that is given, the probabilities of all events are fixed.
- Hence whether or not two events are independent is a matter of 'calculation'

- If A and B are independent then so are A and B^c.
- Using $A = AB + AB^c$, and $AB \subset A$, we have

$$P(AB^c) = P(A-AB) = P(A)-P(AB) = P(A)(1-P(B)) = P(A)P(B^c)$$

- This also shows that A^c and B are independent and so are A^c and B^c.
- For example, in the previous problem, once we saw that F and C are independent, we can conclude M and C are also independent (because in this example we are taking $F^c = M$).

- In many situations calculating probabilities of intersection of events is difficult.
- One often **assumes** A and B are independent to calculate $P(AB)$.
- As we saw, if A and B are independent, then $P(A|B) = P(A)$
- This is often used, at an intuitive level, to justify assumption of independence.

- Consider the example of three tosses of a coin
- Assuming outcomes are equally likely is fine if coin is fair.
- How should we assign these if coin is biased?
- We can assume tosses are independent.
  Then we get, e.g.,
  $P(HTH) = p(1-p)p$ where $p = P(H)$.
- (For a fair coin "equally likely" implies independence and vice versa)

Independence of multiple events

- Events $A_1, A_2, \cdots, A_n$ are said to be (totally) independent if for any k, $1 \leq k \leq n$, and any indices $i_1, \cdots, i_k$, we have

$$P(A_{i_1} \cdots A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$$

- For example, A, B, C are independent if

$$P(AB) = P(A)P(B);\ P(AC) = P(A)P(C);$$

$$P(BC) = P(B)P(C);\ P(ABC) = P(A)P(B)P(C)$$

Pair-wise independence

- Events $A_1, A_2, \cdots, A_n$ are said to be pair-wise independent if

$$P(A_i A_j) = P(A_i)P(A_j), \ \forall i \neq j$$

- Events may be pair-wise independent but not (totally) independent.
- Example: Four balls in a box inscribed with '1', '2', '3' and '123'. Let E_i be the event that number 'i' appears on a radomly drawn ball, $i = 1, 2, 3$.
- Easy to see: $P(E_i) = 0.5, \ i = 1, 2, 3$.
- $P(E_i E_j) = 0.25 \ (i \neq j) \Rightarrow$ pairwise independent
- But, $P(E_1 E_2 E_3) = 0.25 \neq (0.5)^3$

Conditional Independence

- Events A, B are said to be (conditionally) independent given C if
$$P(AB|C) = P(A|C)P(B|C)$$
- If the above holds
$$P(A|BC) = \frac{P(ABC)}{P(BC)} = \frac{P(AB|C)P(C)}{P(BC)}$$
$$= \frac{P(A|C)\,P(B|C)P(C)}{P(BC)} = P(A|C)$$
- Events may be conditionally independent but not independent. (e.g., 'independent' multiple tests for confirming a disease)
- It is also possible that A, B are independent but are not conditionally independent given some other event C.

Use of conditional independence in Bayes rule

- We can write Bayes rule with multiple conditioning events.

$$P(A|BC) = \frac{P(BC|A)P(A)}{P(BC|A)P(A) + P(BC|A^c)P(A^c)}$$

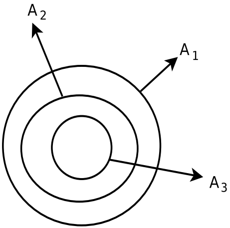
- The above gets simplified if we assume $P(BC|A) = P(B|A)P(C|A)$, $P(BC|A^c) = P(B|A^c)P(C|A^c)$
- Consider the old example, where now we repeat the test for the disease.
- Take: $A = D$, $B = T_+^1$, $C = T_+^2$.
- Assuming conditional independence we can calculate the new posterior probability using the same information we had about true positive and false positive rate.

- We next look at limits for sequences of events.
- A sequence of sets, $A_1, A_2, \cdots$, is said to be monotone decreasing if

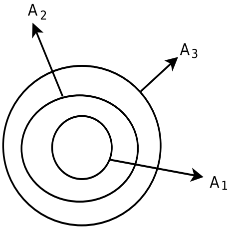
$$A_{n+1} \subset A_n, \ \forall n \quad \text{(denoted as } A_n \downarrow\text{)}$$

- A sequence, $A_1, A_2, \cdots$, is said to be monotone increasing if

$$A_n \subset A_{n+1}, \ \forall n \quad \text{(denoted as } A_n \uparrow\text{)}$$



$A_n \downarrow$



$A_n \uparrow$

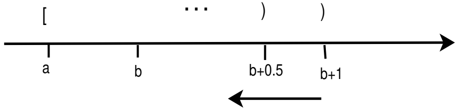
- Let $A_n \downarrow$. Then we define its limit as

$$\lim_{n\to\infty} A_n = \cap_{k=1}^{\infty} A_k$$

- This is reasonable because, when $A_n \downarrow$, we have $A_n \subset A_{n-1} \subset A_{n-2} \cdots$ and hence, $A_n = \cap_{k=1}^{n} A_k$.
- Similarly, when $A_n \uparrow$, we define the limit as

$$\lim_{n\to\infty} A_n = \cup_{k=1}^{\infty} A_k$$

- Let us look at simple examples of monotone sequences of subsets of $\Re$.
- Consider a sequence of intervals:
  $A_n = [a,\ b + \frac{1}{n}), n = 1, 2, \cdots$ with $a, b \in \Re,\ a < b$.



- We have $A_n \downarrow$ and $\lim A_n = \cap_i A_i = [a,\ b]$
- Why? – because
  - $b \in A_n, \forall n \Rightarrow b \in \cap_i A_i$, and
  - $\forall \epsilon > 0,\ b + \epsilon \notin A_n$ after some n (when $\frac{1}{n} < \epsilon$)
    $\Rightarrow b + \epsilon \notin \cap_i A_i$.
    For example, $b + 0.01 \notin A_{101} = [a,\ b + \frac{1}{101})$.

Continuity properties of Probability

- To summarize, limits of monotone sequences of events are defined as follows

$$A_n \downarrow \quad \lim_{n\to\infty} A_n = \cap_{k=1}^{\infty} A_k$$

$$A_n \uparrow \quad \lim_{n\to\infty} A_n = \cup_{k=1}^{\infty} A_k$$

- One can show that

$$P\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} P(A_n)$$

  when the sequence is monotone.
- Known as monotone sequential continuity of probability

Random Variables

- A random variable (on a probability space $(\Omega, \mathcal{F}, P)$) is a real-valued function, $X : \Omega \to \Re$
- For example, $\Omega = \{H, T\}$, $X(H) = 1, \; X(T) = 0$.
- Any random variable results in a new probability space:

$$(\Omega, \mathcal{F}, P) \; \stackrel{X}{\to} \; (\Re, \mathcal{B}, P_X)$$

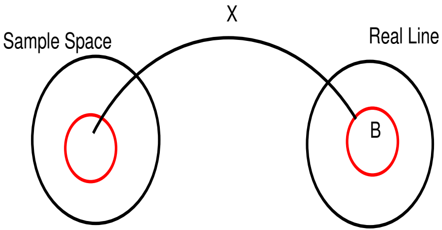
  where $\Re$ is the new sample space and $\mathcal{B} \subset 2^{\Re}$ is the new set of events and P_X is a probability (on $\mathcal{B}$).
- Given $B \subset \Re$, $B \in \mathcal{B}$, we need to know how to get $P_X(B)$.

- Given a probability space $(\Omega, \mathcal{F}, P)$ and a random variable X

$$(\Omega, \mathcal{F}, P) \stackrel{X}{\rightarrow} (\Re, \mathcal{B}, P_X)$$

we define P_X:

$$P_X(B) = P\left(\{\omega \in \Omega \ : \ X(\omega) \in B\}\right), \ B \in \mathcal{B}$$



- We defined P_X:

$$P_X(B) = P\left(\{\omega \in \Omega \ : \ X(\omega) \in B\}\right), \ B \in \mathcal{B}$$

- We use the notation

$$[X \in B] = \{\omega \in \Omega \ : \ X(\omega) \in B\}$$

- So, now we can write

$$P_X(B) = P([X \in B]) = P[X \in B]$$

- We can easily verify P_X is a probability. It satisfies the axioms.
- For the definition of P_X to be proper, for each $B \in \mathcal{B}$, we must have $[X \in B] \in \mathcal{F}$.
  We will assume that.

- A random variable defined on $(\Omega, \mathcal{F}, P)$ results in a new or induced probability space $(\Re, \mathcal{B}, P_X)$.
- Thus, we can study probability models by taking $\Re$ as sample space through the use of random variables.
- Here events are subsets of $\Re$.
- Because of some technical issues we can NOT take $\mathcal{B} = 2^{\Re}$.
- We take $\mathcal{B}$ to be set of all so called Borel sets.
- All intervals (including singleton sets), all sets that can be obtained using countable unions, intersections and complements of intervals are all Borel sets.
- $\mathcal{B}$ is closed under complements, countable unions and intersections (by definition)
- Hence called Borel σ-algebra.

A simple example

- Let $\Omega = \{H, T\}^3 = \{HHH, HHT, \cdots, TTT\}$.
  Let P be specified through 'equally likely' assignment.
  Let $X(\omega)$ be number of H's in ω. Thus, $X(THT) = 1$.
  (X takes one of the values: 0, 1, 2, or 3)
- We can write down $[X \in B]$ for different $B \subset \Re$

$$[X \in (0,\ 1]\,] = \{\omega \in \Omega : X(\omega) \in (0, 1]\} = \{HTT, THT, TTH\};$$

$$[X \in (-1.2,\ 2.78)\,] = \Omega - \{HHH\}$$

- Hence

$$P_X((0,\ 1]) = \frac{3}{8};\ P_X((-1.2,\ 2.78)) = \frac{7}{8}$$

Distribution function of a random variable

- Let X be a random variable on $(\Omega, \mathcal{F}, P)$.
  The (cumulative) distribution function of X is:
  $F_X : \Re \rightarrow \Re$ defined by

  $$F_X(x) = P(\{\omega \in \Omega \ : \ X(\omega) \leq x\}) = P[X \in (-\infty, x]]$$

  We write the event $\{\omega \ : \ X(\omega) \leq x\}$ as $[X \leq x]$.
  We follow this notation with any such relation statement involving X
  e.g., $[X \neq 3]$ represents the event $\{\omega \in \Omega \ : \ X(\omega) \neq 3\}$.
- Thus we have

$$F_X(x) = P[X \leq x] = P(\{\omega \in \Omega \ : \ X(\omega) \leq x\}) = P_X(\,(-\infty,\ x]\,)$$

- The df, F_X, completely specifies the P_X.
  That is, if we know F_X then we can (in principle) compute probability of any $B \in \mathcal{B}$.

Properties of Distribution Functions

- The distribution function of random variable X is given by

$$F_X(x) = P[X \leq x] \ \ (= P(\{\omega \ : \ X(\omega) \leq x\}))$$

- Any distribution function should satisfy the following:
  1. $0 \leq F_X(x) \leq 1, \ \forall x$
  2. $F_X(-\infty) = 0; \ F_X(\infty) = 1$
  3. F_X is non-decreasing: $x_1 \leq x_2 \ \Rightarrow \ F_X(x_1) \leq F_X(x_2)$
     This is because

$$x_1 \leq x_2 \ \Rightarrow \ [X \leq x_1] \subseteq [X \leq x_2] \ \Rightarrow \ \ F_X(x_1) \leq F_X(x_2)$$

  4. F_X is right continuous and has left-hand limits.

Right Continuity and left limits

We have

$$\lim_{x_n \downarrow x} (-\infty, x_n] = (-\infty, x], \quad \lim_{x_n \uparrow x} (-\infty, x_n] = (-\infty, x)$$

$$\lim_{x_n \downarrow x} P_X((-\infty, x_n]) = P_X((-\infty, x]), \quad \lim_{x_n \uparrow x} P_X((-\infty, x_n]) = P_X((-\infty, x))$$

Hence

$$F_X(x^+) = \lim_{x_n \downarrow x} F_X(x_n) = F_X(x) = P_X(\ (-\infty,\ x]\ )$$

$$F_X(x^-) = \lim_{x_n \uparrow x} F_X(x_n) = P_X(\ (-\infty,\ x)\ )$$

- If $A \subset B$ then $P(B - A) = P(B) - P(A)$
- We have $(-\infty,\ x] - (-\infty,\ x) = \{x\}$. Hence

$$P_X(\ (-\infty,\ x]\ ) - P_X(\ (-\infty,\ x)\ ) = P_X(\{x\}) = P(\{\omega\ :\ X(\omega) = x\})$$

- Thus we get

$$F_X(x^+) - F_X(x^-) = P[X = x]\ \ (= P(\{\omega\ :\ X(\omega) = x\}))$$

- When F_X is discontinuous at x the height of discontinuity is the probability that X takes that value.
- And, if F_X is continuous at x then $P[X = x] = 0$

Distribution Functions

- Let X be a random variable.
- Its distribution function, $F_X : \Re \rightarrow \Re$ is given by $F_X(x) = P[X \leq x]$
- The distribution function satisfies
  1. $0 \leq F_X(x) \leq 1, \ \forall x$
  2. $F_X(-\infty) = 0; \ F_X(\infty) = 1$
  3. F_X is non-decreasing: $x_1 \leq x_2 \ \Rightarrow \ F_X(x_1) \leq F_X(x_2)$
  4. F_X is right continuous and has left-hand limits.
- We also have $F_X(x^+) - F_X(x^-) = P[X = x]$

- $F_X(x) = P[X \leq x] = P[X \in (-\infty,\ x]\ ]$
- Given F_X, we can, in principle, find $P[X \in B]$ for all Borel sets.
- In particular, for $a < b$,

$$
\begin{aligned}
P[a < X \leq b] &= P[X \in (a,\ b]\ ] \\
&= P[X \in (\ (-\infty,\ b] - (-\infty,\ a]\ )\ ] \\
&= P[X \in (-\infty,\ b]\ ] - P[X \in (-\infty,\ a]\ ] \\
&= F_X(b) - F_X(a)
\end{aligned}
$$

- $P[a \leq X \leq b] = F_X(b) - F_X(a) + P[X = a]$

- There are two classes of random variables that we would study here.
- These are called discrete and continuous random variables.
- There can be random variables that are neither discrete nor continuous.
- But these two are important classes of random variables.
- Note that the distribution function is defined for **all** random variables.

Discrete Random Variables

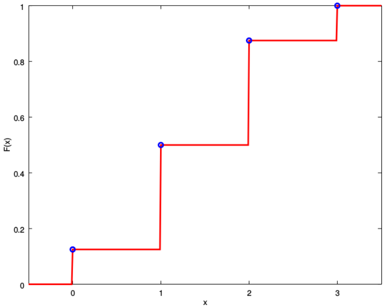
- A random variable X is said to be discrete if it takes only countably many distinct values.
- Countably many means finite or countably infinite.
- Any random variable defined on a countable Ω would be discrete.

Discrete Random Variable Example

- Consider three independent tosses of a fair coin.
- $\Omega = \{H,\ T\}^3$ and $X(\omega)$ is the number of H's in ω.
- This rv takes four distinct values, namely, $0, 1, 2, 3$.
- We denote this as $X \in \{0, 1, 2, 3\}$

- Let X be a discrete rv with $X \in \{x_1, x_2, \cdots\}$.
  (As a notation we assume $x_1 < x_2 < \cdots$).
- Let $q_i = P[X = x_i]$ ($q_i \geq 0$ and $\sum_i q_i = 1$).
- We know that $F_X(x) - F_X(x^-) = P[X = x]$.
- Thus the distribution function would be a stair-case function with jumps of magnitude q_i at each x_i.
- The distribution function of X is specified completely by these q_i
  (Here we are assuming that the intervals (x_i, x_{i+1}) are non-empty).

- Consider example of tossing three coins where X is number of heads.
  Here $X \in \{0, 1, 2, 3\}$.
- The plot of its distribution function is:



- This is a stair-case function.
- It has jumps at $x = 0, 1, 2, 3$, which are the values that X takes. In between these it is constant.
- The jump at, e.g., $x = 2$ is $3/8$ which is the probability of X taking that value.

probability mass function, f_X

- Let X be a discrete rv with $X \in \{x_1, x_2, \cdots\}$.
- The probability mass function (pmf) of X is defined by

$$f_X(x_i) = P[X = x_i]; \quad f_X(x) = 0, \quad \text{for all other } x$$

- f_X is also a real-valued function of a real variable.
- We can write the definition compactly as $f_X(x) = P[X = x]$
- The distribution function (df) and the pmf are related as

$$f_X(x) = F_X(x) - F_X(x^-)$$

$$F_X(x) = P[X \leq x] = \sum_{i : x_i \leq x} f_X(x_i)$$

Properties of pmf

- The probability mass function of a discrete random variable $X \in \{x_1, x_2, \cdots\}$ satisfies
  1. $f_X(x) \geq 0, \forall x$ and $f_X(x) = 0$ if $x \neq x_i$ for some i
  2. $\sum_i f_X(x_i) = 1$
- Any function satisfying the above two would be a pmf of some discrete random variable.
- We can specify a discrete random variable by giving either F_X or f_X.
- Distribution function is defined for any random variable. But pmf is defined only for discrete random variables

- Any discrete random variable can be specified by
  - giving the set of values of X, $\{x_1, x_2, \cdots\}$, and
  - numbers q_i such that $q_i = P[X = x_i] = f_X(x_i)$
- Note that we must have $q_i \geq 0$ and $\sum_i q_i = 1$.
- As we saw this is how we can specify a probability assignment on any countable sample space.
- Any random variable on a countable sample space would be discrete.

Computations of Probabilities for discrete rv's

- A discrete random variable is specified by giving either df or pmf. One can be obtained from the other.
- We normally specify it through the pmf.
- Given $X \in \{x_1, x_2, \cdots\}$ and f_X, we can (in principle) compute probability of any event

$$P[X \in B] = \sum_{\substack{i: \\ x_i \in B}} f_X(x_i)$$

- For example, if $X \in \{0, 1, 2, 3\}$ then

$$P[X \in [0.5,\ 1.32] \cup [2.75,\ 5.2]\ ] = f_X(1) + f_X(3)$$

- We next look at some standard discrete random variable models

Bernoulli Distribution

- Bernoulli random variable: $X \in \{0, 1\}$ with

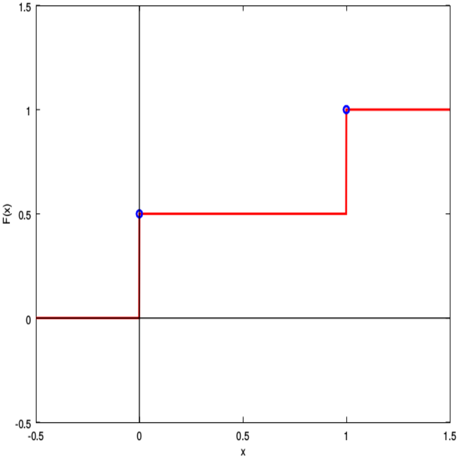
  $f_X(1) = p$; $f_X(0) = 1-p$; where $0 < p < 1$ is a parameter

- This f_X is easily seen to be a pmf
- Consider $(\Omega, \mathcal{F}, P)$ with $B \in \mathcal{F}$. (The Ω here may be uncountable).
- Consider the random variable

$$I_B(\omega) = \begin{cases} 0 & \text{if } \omega \notin B \\ 1 & \text{if } \omega \in B \end{cases}$$

- It is called indicator (random variable) of B.
- $P[I_B = 1] = P(\{\omega \ : \ I_B(\omega) = 1\}) = P(B)$
- Thus, this indicator rv has Bernoulli distribution with $p = P(B)$

The df of a Bernoulli rv



Binomial Distribution

- $X \in \{0,\ 1,\ \cdots,\ n\}$ with pmf

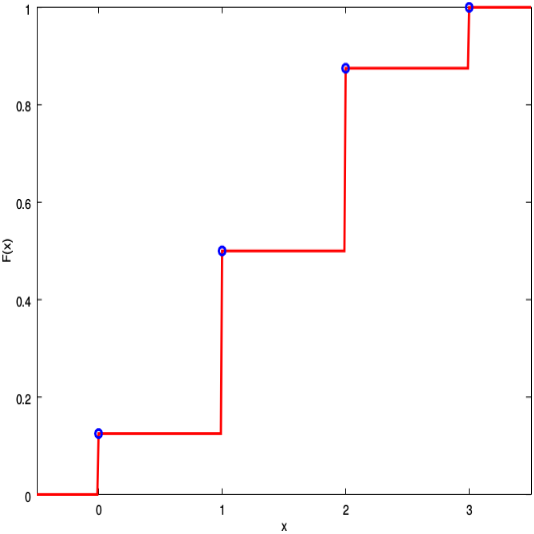
$$f_X(k) = \ {}^nC_k\ p^k\ (1-p)^{n-k},\ k = 0, 1, \cdots, n$$

  where n, p are parameters (n is a +ve integer and $0 < p < 1$).
- This is easily seen to be a pmf

$$\sum_{k=0}^{n} {}^nC_k\ p^k\ (1-p)^{n-k} = (p + 1 - p)^n = 1$$

- Consider n independent tosses of coin whose probability of heads is p. If X is the number of heads then X has the above binomial distribution.
  (Number of successes in n bernoulli trials)

The example we considered was that of Binomial



Poisson Distribution

- $X \in \{0, \ 1, \ 2, \cdots\}$ with pmf
$$f_X(k) = \frac{\lambda^k \ e^{-\lambda}}{k!}, \ k = 0, 1, 2, \cdots$$
where $\lambda > 0$ is a parameter.
- We can see this to be a pmf by
$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \ e^{-\lambda} = e^{\lambda} \ e^{-\lambda} = 1$$
- Poisson distribution is also useful in many applications

Geometric Distribution

- $X \in \{1,\ 2, \cdots\}$ with pmf

$$f_X(k) = (1-p)^{k-1}\ p,\ \ k = 1, 2, \cdots$$

  where $0 < p < 1$ is a parameter.
- Consider tossing a coin (with prob of H being p) repeatedly till we get a head. X is the toss number on which we got the first head.
- In general waiting for 'success' in independent Bernoulli trials.

df of geometric rv

- Suppose X is a geometric rv. Let n be a positive integer.
- Then

$$
\begin{aligned}
P[X > n] &= \sum_{k=n+1}^{\infty} P[X = k] = \sum_{k=n+1}^{\infty} (1-p)^{k-1}\, p \\
&= p\, \frac{(1-p)^n}{1-(1-p)} = (1-p)^n
\end{aligned}
$$

- $F_X(n) = P[X \leq n] = 1 - (1-p)^n$, n a positive integer.
- What is $F_X(x)$ when x is not a positive integer?

Memoryless property of geometric rv

- Let m, n be positive integers. Then

$$\begin{aligned} P[X > m+n | X > m] &= \frac{P[X > m+n, X > m]}{P[X > m]} \\ &= \frac{P[X > m+n]}{P[X > m]} \\ &= \frac{(1-p)^{m+n}}{(1-p)^m} = (1-p)^n \\ \Rightarrow P[X > m+n | X > m] &= P[X > n] \end{aligned}$$

- This is known as the memoryless property of geometric distribution
- Same as

$$P[X > m+n] = P[X > m] P[X > n]$$

Memoryless property defines geometric rv

- Suppose $X \in \{0, 1, \cdots\}$ is a discrete rv satisfying, for all non-negative integers, m, n

$$P[X > m + n] = P[X > m]P[X > n]$$

- Then we can show that X has geometric distribution.

Continuous Random Variables

- A rv, X, is said to be continuous (or of continuous type) if there exists a function $f_X : \Re \rightarrow \Re$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt, \quad \forall x$$

- The f_X is called the probability density function (pdf) of X.
- If X is a continuous rv, then, by definition, F_X is continuous at every x.
- By the fundamental theorem of calculus, we have

$$\frac{dF_X(x)}{dx} = f_X(x), \ \forall x \text{ where } f_X \text{ is continuous}$$

Continuous Random Variables

- If X is a continuous rv then its distribution function, F_X, is continuous.
- Hence a discrete random variable is not a continuous rv!
- If a rv takes countably many values then it is discrete.
- Hence continuous random variables take uncountably many values.
- However, if a rv takes uncoutably infinitely many distinct values, it does not necessarily imply it is of continuous type.
- As mentioned earlier, there would be many random variables that are neither discrete nor continuous.

Continuous Random Variables

- The df of a continuous rv is continuous.
- This implies
  $F_X(x) = F_X(x^+) = F_X(x^-)$
- Hence, if X is a continuous random variable then

$$P[X = x] = F_X(x) - F_X(x^-) = 0, \ \forall x$$

Properties of pdf

- The pdf, $f_X : \Re \to \Re$, of a continuous rv satisfies
  A1. $f_X(x) \geq 0, \;\; \forall x$
  A2. $\int_{-\infty}^{\infty} f_X(t)\, dt = 1$
- Any f_X that satisfies the above two would be the probability density function of a continuous rv
- Given f_X satifying the above two, define

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt, \;\; \forall x$$

  This F_X satisfies
  1. $F_X(-\infty) = 0;\; F_X(\infty) = 1$
  2. F_X is non decreasing.
  3. F_X is continuous (and hence right continuous with left limits)
- This shows the the F_X is a df and hence f_X is a pdf

- Let X be a continuous rv.
- It can be specified by giving either F_X or the pdf, f_X.
- We can, in principle, compute probability of any event as

$$P[X \in B] = \int_B f_X(t)\, dt, \;\; \forall B \in \mathcal{B}$$

- In particular, we have

$$P[X \in [a,\; b]] = P[a \le X \le b] = \int_a^b f_X(t)\, dt = F_X(b) - F_X(a)$$

- For a continuous random variable, X, since $P[X = x] = 0, \forall x$,

$$P[a \le X \le b] = P[a < X \le b] = P[a \le X < b] \text{ etc.}$$

- Recall that for a general rv

$$F_X(b) - F_X(a) = P[a < X \le b]$$

- If X is a continuous rv, we have

$$P[a \leq X \leq b] = \int_a^b f_X(t)\, dt$$

- Thus

$$P[x \leq X \leq x + \Delta x] = \int_x^{x+\Delta x} f_X(t)\, dt \approx f_X(x)\, \Delta x$$

- That is why f_X is called probability density function.

- For any random variable, the df is defined and it is given by
$$F_X(x) = P[X \leq x]$$
- The value of $F_X(x)$ at any x is probability of some event.
- The pmf is defined only for discrete random variables as $f_X(x) = P[X = x]$
- The value of pmf is also a probability
- We use the same symbol for pdf (as for pmf), defined by
$$F_X(x) = \int_{-\infty}^{x} f_X(x)\, dx$$
- Note that the value of pdf is not a probability.
- We can say $f_X(x)\, dx \approx P[x \leq X \leq x + dx]$

- A continuous random variable is a probability model on uncountably infinite Ω.
- For this, we take $\Re$ as our sample space.
- We can specify a continuous rv either through the df or through the pdf.
- We next consider a few standard continuous random variables.

Uniform distribution

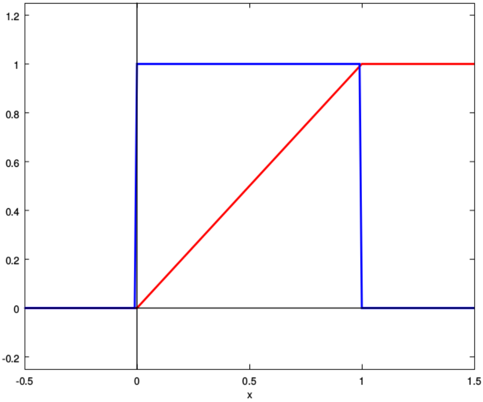
- X is uniform over $[a,\ b]$ when its pdf is

$$f_X(x) = \frac{1}{b-a}, \quad a \le x \le b$$

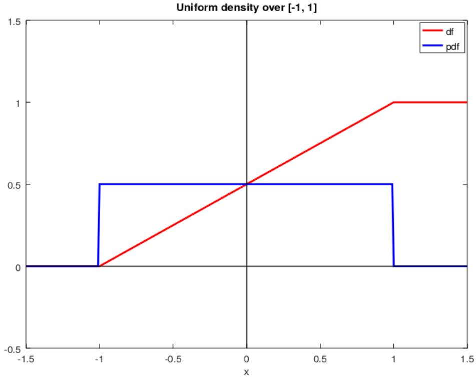
  ($f_X(x) = 0$ for all other values of x).
- Uniform distribution over open or closed interval is essentially the same.
- When X has this distribution, we say $X \sim U[a,\ b]$
- By integrating the above, we can get the df

Uniform density over [0, 1]

- If $X \sim U[0,\ 1]$ then $f_X(x) = 1,\ 0 \leq x \leq 1$.
  $F_X(x) = x,\ 0 < x < 1,\ F_X(x) = 0,\ x \leq 0,$
  $F_X(x) = 1,\ x \geq 1.$



- A plot of density and distribution functions of another uniform rv



- Let $X \sim U[a,\ b]$. Then $f_X(x) = \frac{1}{b-a},\ a \leq x \leq b$
- Let $[c,\ d] \subset [a,\ b]$.
- Then $P[X \in [c,\ d]] = \int_c^d f_X(t)\ dt = \frac{d-c}{b-a}$
- Probability of an interval is proportional to its length.
- Thus this is the analogue of "equally likely"

Exponential distribution

- The pdf of exponential distribution is

$$f_X(x) = \lambda\, e^{-\lambda x}, \;\; x > 0, \; (\lambda > 0 \text{ is a parameter})$$

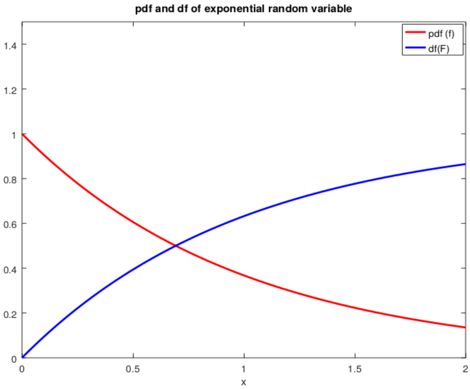
(By our notation, $f_X(x) = 0$ for $x \leq 0$)

- It is easy to verify $\int_0^\infty f_X(x)\, dx = 1$.
- here $F_X(x) = 0$, for $x \leq 0$.
- For $x > 0$ we can compute F_X by integrating f_X:

$$F_X(x) = \int_0^x \lambda\, e^{-\lambda x}\, dx = \lambda \left. \frac{e^{-\lambda x}}{-\lambda} \right|_0^x = 1 - e^{-\lambda x}$$

- This also gives us: $P[X > x] = 1 - F_X(x) = e^{-\lambda x}$ for $x > 0$.

- A plot of density and distribution functions of an exponential rv is given below



exponential distribution is memoryless

- If X has exponential distribution, then, for $t, s > 0$,

$$P[X > t+s] = e^{-\lambda(t+s)} = e^{-\lambda t} \, e^{-\lambda s} = P[X > t] \, P[X > s]$$

- This gives us the memoryless property

$$P[X > t + s \mid X > t] = \frac{[P[X > t + s]}{P[X > t]} = P[X > s]$$

- Exponential distribution is a useful model for, e.g., life-time of components.
- If the distribution of a non-negative continuous random variable is memory less then it must be exponential.

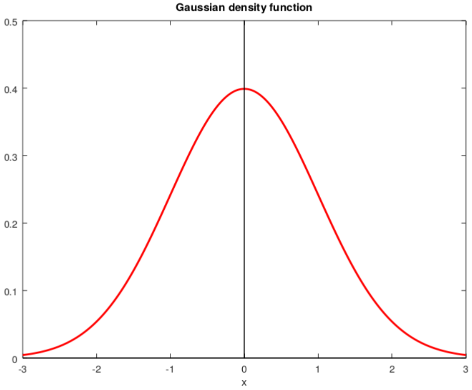
Gaussian Distribution

- The pdf of Gaussian distribution is given by

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

  where $\sigma > 0$ and $\mu \in \Re$ are parameters.
- We write $X \sim \mathcal{N}(\mu, \sigma^2)$ to denote that X has Gaussian density with parameters μ and σ.
- This is also called the Normal distribution.
- The special case where $\mu = 0$ and $\sigma^2 = 1$ is called standard Gaussian (or standard Normal) distribution.

- A plot of Gaussian density functions is given below



- $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \; -\infty < x < \infty$
- Showing that the density integrates to 1 is not trivial.
- Take $\mu = 0, \sigma = 1$. Let $I = \int_{-\infty}^{\infty} f_X(x) \; dx$. Then

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-0.5x^2} \; dx \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-0.5y^2} \; dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} \; e^{-0.5(x^2+y^2)} \; dx \; dy \end{aligned}$$

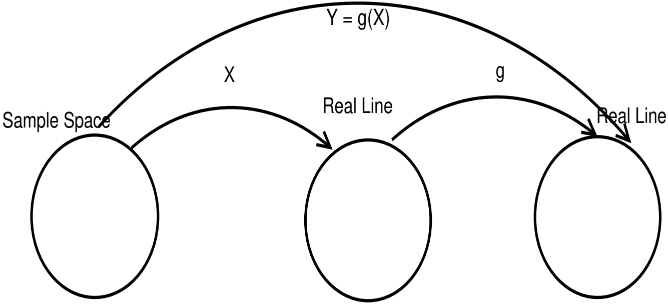
- Now converting the above integral into polar coordinates would allow you to show $I = 1$.

Functions of a random variable

- We next look at random variables defined in terms of other random variables.

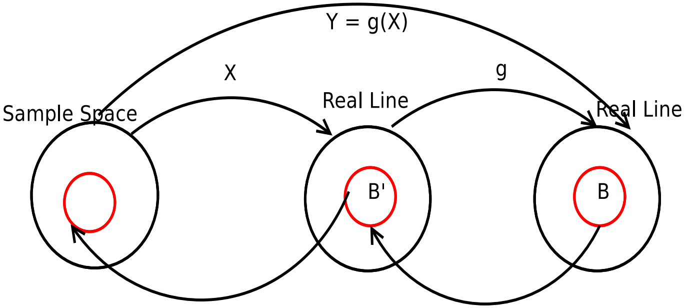
Functions of a Random Variable

- Let X be a rv on some probability space $(\Omega, \mathcal{F}, P)$. (Recall $X : \Omega \to \Re$)
- Consider a function $g : \Re \to \Re$
- Let $Y = g(X)$. Then Y also maps Ω into real line.



Functions of a Random Variable

- Let X be a rv on some probability space $(\Omega, \mathcal{F}, P)$. (Recall $X : \Omega \to \Re$)
- Consider a function $g : \Re \to \Re$
- Let $Y = g(X)$. Then Y also maps Ω into real line.



- If g is a 'nice' function, Y would also be a random variable

- Let X be a rv and let $Y = g(X)$.
- The distribution function of Y is given by

$$\begin{aligned} F_Y(y) &= P[Y \leq y] \\ &= P[g(X) \leq y] \\ &= P[X \in \{z \,:\, g(z) \leq y\}] \end{aligned}$$

- This probability can be obtained from distribution of X.
- Thus, in principle, we can find the distribution of Y if we know that of X

Example

- Let $Y = aX + b, \; a > 0$.
- Then we have

$$\begin{aligned} F_Y(y) &= P[Y \leq y] \\ &= P[aX + b \leq y] \\ &= P[aX \leq y - b] \\ &= P\left[X \leq \frac{y-b}{a}\right], \quad \text{since } a > 0 \\ &= F_X\left(\frac{y-b}{a}\right) \end{aligned}$$

- This tells us how to find df of Y when it is an affine function of X.
- If X is continuous rv, then, $f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right)$

- Let $X \sim \mathcal{N}(0,1)$. That is, $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
- Let $Y = aX + b$. Then the pdf of Y is

$$\begin{aligned} f_Y(y) &= \frac{1}{a} f_X\left(\frac{y-b}{a}\right) \\ &= \frac{1}{a\sqrt{2\pi}} e^{-\frac{(y-b)^2}{2a^2}} \end{aligned}$$

- This shows that $Y \sim \mathcal{N}(b, a^2)$
  Linear transformation of a Gaussian is a Gaussian.
- Similarly you can show that if X is uniform over $[0,1]$ and $Y = aX + b$ then Y is uniform over $[b, a+b]$ (assuming $a > 0$).

- Suppose X is a discrete rv with $X \in \{x_1, x_2, \cdots\}$.
- Suppose $Y = g(X)$.
- Then Y is also discrete and Y takes values $g(x_i)$.
- We can find the pmf of Y as

$$\begin{aligned} f_Y(y) &= p[Y = y] = P[g(X) = y] \\ &= P[X \in \{x_i \ : \ g(x_i) = y\}] \\ &= \sum_{\substack{i: \\ g(x_i)=y}} f_X(x_i) \end{aligned}$$

- Let $Y = X^2$.
- For $y < 0$, $F_Y(y) = P[Y \leq y] = 0$ (since $Y \geq 0$)
- For $y \geq 0$, we can get $F_Y(y)$ as

$$
\begin{aligned}
F_Y(y) &= P[Y \leq y] = P[X^2 \leq y] \\
&= P[-\sqrt{y} \leq X \leq \sqrt{y}] \\
&= P[-\sqrt{y} < X \leq \sqrt{y}] + P[X = -\sqrt{y}] \\
&= F_X(\sqrt{y}) - F_X(-\sqrt{y}) + P[X = -\sqrt{y}]
\end{aligned}
$$

- If X is a continuous random variable, then we get

$$
\begin{aligned}
f_Y(y) &= \frac{d}{dy} \left(F_X(\sqrt{y}) - F_X(-\sqrt{y})\right) \\
&= \frac{1}{2\sqrt{y}} \left[f_X(\sqrt{y}) + f_X(-\sqrt{y})\right]
\end{aligned}
$$

- This is the general formula for density of X^2 when X is continuous rv.

- Let $X \sim \mathcal{N}(0,1)$: $f_X(x) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{x^2}{2}}$
- Let $Y = X^2$. Then we know $f_Y(y) = 0$ for $y < 0$. For $y \geq 0$,

$$
\begin{aligned}
f_Y(y) &= \frac{1}{2\sqrt{y}}\left[f_X(\sqrt{y}) + f_X(-\sqrt{y})\right] \\
&= \frac{1}{2\sqrt{y}}\left[\frac{1}{\sqrt{2\pi}}\ e^{-\frac{y}{2}} + \frac{1}{\sqrt{2\pi}}\ e^{-\frac{y}{2}}\right] \\
&= \frac{1}{2\sqrt{y}}\frac{2}{\sqrt{2\pi}}\ e^{-\frac{y}{2}} \\
&= \frac{1}{\sqrt{\pi}}\ \left(\frac{1}{2}\right)^{0.5}\ y^{-0.5}\ e^{-\frac{1}{2}y}
\end{aligned}
$$

- This is an example of gamma density.

Gamma density

- The Gamma function is given by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\, e^{-x}\, dx, \;\; \alpha > 0$$

It can be easily verified that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$.

- The Gamma density is given by

$$f(x) = \frac{1}{\Gamma(\alpha)}\, \lambda^\alpha x^{\alpha-1} e^{-\lambda x} = \frac{1}{\Gamma(\alpha)}\, (\lambda x)^{\alpha-1}\, \lambda e^{-\lambda x}, \;\; x > 0$$

Here $\alpha, \lambda > 0$ are parameters.

- The earlier density we saw corresponds to $\alpha = \lambda = 0.5$:

$$f_Y(y) = \frac{1}{\sqrt{\pi}}\, \left(\frac{1}{2}\right)^{0.5}\, y^{-0.5}\, e^{-\frac{1}{2}y}, \;\; y > 0$$

- The gamma density with parameters $\alpha, \lambda > 0$ is given by

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha - 1} e^{-\lambda x}, \quad x > 0$$

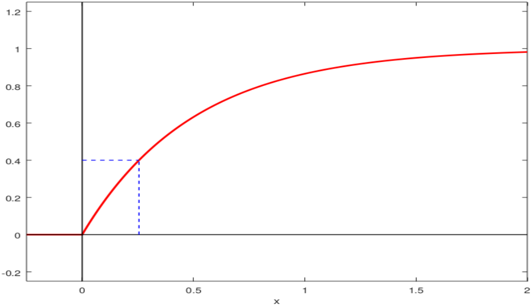
- If $X \sim \mathcal{N}(0, 1)$ then X^2 has gamma density with parameters $\alpha = \lambda = 0.5$.
- When α is a positive integer then the gamma density is known as the Erlang density.
- If $\alpha = 1$, gamma density becomes exponential density.
- If $\lambda = 0.5$ and $\alpha = \frac{n}{2}$ (where n is a positive integer) then the Gamma density is called chi-square density with n degrees of freedom.

- Let G be a continuous invertible distribution function.
- Let $X \sim U[0,\ 1]$ and let $Y = G^{-1}(X)$.
- We can get the df of Y as

$$F_Y(y) = P[Y \leq y] = P[G^{-1}(X) \leq y] = P[X \leq G(y)] = G(y)$$

- Thus, starting with uniform rv, we can generate a rv with a desired distribution.
- Very useful in random number generation. Known as the inverse function method.
- Can be generalized to handle any df. It only involves defining an 'inverse' suitably. (Left as an exercise!)

- We can visualize this as shown below



- Suppose we want to simulate exponential rv
- The df is $F(x) = 1 - e^{-\lambda x}$. We can invert this.

$$y = 1 - e^{-\lambda x} \;\Rightarrow\; e^{-\lambda x} = 1 - y \;\Rightarrow x = -\frac{1}{\lambda} \ln(1 - y)$$

- Hence, if $X \sim U[0,\ 1]$, then $Y = \frac{-1}{\lambda} \ln(1 - X)$ would be exponential

- Let X be a cont rv with an invertible distribution function, say, F.
- Define $Y = F(X)$.
- Since range of F is $[0, \ 1]$, we know $0 \leq Y \leq 1$.
- For $0 \leq y \leq 1$ we can obtain $F_Y(y)$ as

$$F_Y(y) = P[Y \leq y] = P[F(X) \leq y] = P[X \leq F^{-1}(y)] = F(F^{-1}(y)) = y$$

- This means Y has uniform density.
- Has interesting applications.
  E.g., histogram equalization in image processing

- Let us sum-up the last two examples
- If $X \sim U[0,\ 1]$ and $Y = F^{-1}(X)$, then Y has df F.
- If df of X is F and $Y = F(X)$ then Y is uniform over $[0,\ 1]$.

A useful theorem

- Let $g : \Re \rightarrow \Re$ be differentiable with $g'(x) > 0, \forall x$ or $g'(x) < 0, \forall x$.
- Let X be a continuous rv and let $Y = g(X)$.
- Then Y is a continuous rv with pdf

$$f_Y(y) = f_X(g^{-1}(y)) \, \left| \frac{d}{dy} g^{-1}(y) \right|, \; a \leq y \leq b$$

  where $a = \min(g(\infty), \; g(-\infty))$ and $b = \max(g(\infty), \; g(-\infty))$
- We omit the proof.

Expectation of a discrete rv

- Let X be a discrete rv with $X \in \{x_1, x_2, \cdots\}$
- We define its expectation by

$$E[X] = \sum_i x_i \; f_X(x_i)$$

- Expectation is essentially a weighted average.

Expectation of a Continuous rv

- If X is a continuous random variable with pdf, f_X, we define its expectation as

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

- Sometimes one uses the following notation to denote expectation of both kinds of rv

$$E[X] = \int_{-\infty}^{\infty} x \, dF_X(x)$$

- Though we consider only discrete or continuous rv's, expectation is defined for all random variables.

$$E[X] = \sum_i x_i f_X(x_i) \ \text{ or } \ \int_{-\infty}^{\infty} x f_X(x)\, dx$$

- $E[X]$ is a real number.
- We some times write EX for $E[X]$

Binary random variable

- Expectation of a binary rv (e.g., Bernoulli):

$$EX = 0 \times f_X(0) + 1 \times f_X(1) = P[X = 1]$$

- Expectation of a binary random variable is same as the probability of the rv taking value 1.
- Thus, for example, $EI_A = P(A)$.

Expectation of Poisson rv

- $f_X(k) = \frac{\lambda^k}{k!}\, e^{-\lambda},\ k = 0, 1, \cdots$

$$EX = \sum_{k=0}^{\infty} k\, \frac{\lambda^k}{k!}\, e^{-\lambda} = \lambda\, e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda\, e^{-\lambda} \sum_{k'=0}^{\infty} \frac{\lambda^{k'}}{k'!} = \lambda$$

Expectation of Geometric rv

- $f_X(k) = (1-p)^{k-1}\, p, \quad k = 1, 2, \cdots$

$$EX = \sum_{k=1}^{\infty} k\, (1-p)^{k-1}\, p$$

- We have

$$\sum_{k=1}^{\infty} (1-p)^k = \frac{1-p}{p} = \frac{1}{p} - 1$$

- Term-wise differentiation of the above gives

$$\sum_{k=1}^{\infty} k\, (1-p)^{k-1} = \frac{1}{p^2}$$

- This gives us $EX = \frac{1}{p}$

Expectation of Binomial rv

- Let $f_X(k) = {}^nC_k \, p^k(1-p)^{n-k}, \; k = 0, 1, \cdots, n.$

$$\begin{aligned} EX &= \sum_{k=0}^{n} k \, \frac{n!}{k!(n-k)!} \, p^k \, (1-p)^{n-k} \\ &= np \end{aligned}$$

Expectation of uniform rv

- Let $X \sim U[a,b]$. $f_X(x) = \frac{1}{b-a},\ a \le x \le b$

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} x\, f_X(x)\, dx \\ &= \int_a^b x\, \frac{1}{b-a}\, dx \\ &= \frac{1}{b-a} \left. \frac{x^2}{2} \right|_a^b \\ &= \frac{1}{b-a}\, \frac{b^2-a^2}{2} \\ &= \frac{b+a}{2} \end{aligned}$$

Expectation of exponential density

- $f_X(x) = \lambda \, e^{-\lambda x}, \;\; x > 0.$

$$\begin{aligned} EX &= \int_0^\infty x \, \lambda \, e^{-\lambda x} \, dx \\ &= x \, \lambda \, \frac{e^{-\lambda x}}{-\lambda} \Big|_0^\infty - \int_0^\infty \lambda \, \frac{e^{-\lambda x}}{-\lambda} \, dx \\ &= \int_0^\infty e^{-\lambda x} \, dx \\ &= \frac{1}{\lambda} \end{aligned}$$

Expectation of Gaussian density

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

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} x\, \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\ &\quad \text{make a change of variable } \ y = \frac{x-\mu}{\sigma} \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, (\sigma y + \mu) e^{-\frac{y^2}{2}}\, dy \\ &= \mu \end{aligned}$$

Expectation of a function of a random variable

- Let X be a rv and let $Y = g(X)$.
- **Theorem**: $EY = \int y \; dF_Y(y) = \int g(x) \; dF_X(x)$
- That is, if X is discrete, then

$$EY = \sum_j y_j \; f_Y(y_j) = \sum_i g(x_i) f_X(x_i)$$

- If X and Y are continuous

$$EY = \int y \; f_Y(y) \; dy = \int g(x) \; f_X(x) \; dx$$

- This theorem is true for all rv's.
  (Some people call it the LOTUS theorm – Law Of The Unconscious Statistician)

Some Properties of Expectation

$$E[g(X)] = \sum_i g(x_i) f_X(x_i) \quad \text{or} \quad E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$$

- If $X \geq 0$ then $EX \geq 0$
- $E[b] = b$ where b is a constant
- $E[ag(X)] = aE[g(X)]$ where a is a constant
- $E[aX + b] = aE[X] + b$ where a, b are constants.
- $E[ag_1(X) + bg_2(X)] = aE[g_1(X)] + bE[g_2(X)]$

- Consider the problem: $\min_c \; E[(X-c)^2]$
- We are asking what is the best constant to approximate a rv with
- We are trying to minimize (weighted) average, over all values X can take, of the square of the error
- We are interested in the best mean-square approximation of X by a constant.

$$E[(X-c)^2] = E[X^2 + c^2 - 2cX] = E[X^2] + c^2 - 2cE[X]$$

- We differentiate this and equate to zero to get the best c
$2c^* = 2E[X] \;\Rightarrow\; c^* = E[X]$
- Thus, $E[(X - EX])^2] \leq E[(X-c)^2], \forall c$.

Variance of a Random variable

- We define variance of X as $E[(X - EX)^2]$ and denote it as $\mathrm{Var}(X)$.
- By definition, $\mathrm{Var}(X) \geq 0$.

$$\begin{aligned} \mathrm{Var}(X) &= E[(X - EX)^2] \\ &= E\left[X^2 + (EX)^2 - 2X(EX)\right] \\ &= E[X^2] + (EX)^2 - 2(EX)E[X] \\ &= E[X^2] - (EX)^2 \end{aligned}$$

- This also implies: $E[X^2] \geq (EX)^2$

Some properties of variance

- $\text{Var}(X + c) = \text{Var}(X)$ where c is a constant

$$\text{Var}(X+c) = E\left[\{(X + c) \; - \; E[X + c]\}^2\right] = E\left[(X - EX)^2\right] = \text{Var}(X)$$

- $\text{Var}(cX) = c^2\text{Var}(X)$ where c is a constant

$$\text{Var}(cX) = E\left[(cX - E[cX])^2\right] = E\left[(cX - cE[X])^2\right] = c^2\text{Var}(X)$$

Variance of uniform rv

- $f_X(x) = \frac{1}{b-a}, \ a \leq x \leq b$

$$\begin{aligned} E[X^2] &= \int_a^b x^2 \, \frac{1}{b-a} \, dx \\ &= \frac{b^2 + ab + a^2}{3} \end{aligned}$$

- Now we get variance as

$$\begin{aligned} \text{Var}(X) &= EX^2 - (EX)^2 = \frac{b^2 + ab + a^2}{3} - \frac{(b+a)^2}{4} \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

- We can use $\text{Var}(X) = E[X^2] - (EX)^2$ to obtain variances of all standard random variables.
- When X is exponential
  $f_X(x) = \lambda e^{-\lambda x}, x > 0, \quad \text{Var}(X) = \frac{1}{\lambda^2}$
- For Binomial and Poisson it is easier to use
  $E[X^2] = E[X(X-1)] + E[X]$
- When X is Binomial
  $f_X(k) = {}^nC_k \; p^k(1-p)^{n-k}, \quad \text{Var}(X) = np(1-p)$
- When X is Poisson
  $f_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad \text{Var}(X) = \lambda$

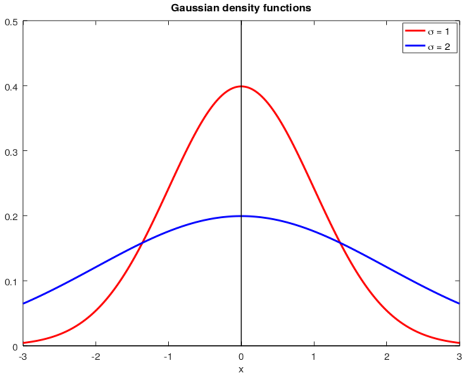
Variance of Gaussian rv

- Let $X \sim \mathcal{N}(0,1)$: $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \; -\infty < x < \infty$, $EX = 0$.
- We know $EX = 0$. Hence $\text{Var}(X) = EX^2$.

$$
\begin{aligned}
\text{Var}(X) &= EX^2 = \int_{-\infty}^{\infty} x^2 \, \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \; dx \\
&= \int_{-\infty}^{\infty} x \, \left(x \, \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \right) \; dx \\
&= x \; \frac{-1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \; dx \\
&= 1
\end{aligned}
$$

- Let $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \ -\infty < x < \infty.$
- We know $EX = 0$ and $\text{Var}(X) = 1$
- Let $Y = \sigma X + \mu$. Then $Y \sim \mathcal{N}(\mu, \sigma^2)$.
- Since $Y = \sigma X + \mu$, we get
  - $EY = \sigma EX + \mu = \mu$
  - $\text{Var}(Y) = \sigma^2 \text{Var}(X) = \sigma^2$
- When $Y \sim \mathcal{N}(\mu, \sigma^2)$, $EY = \mu$ and $\text{Var}(Y) = \sigma^2$.

- Here is a plot of Gaussian densities with different variances



moments of a random variable

- We define the k^{th} order moment of a rv, X, by

$$m_k = E[X^k] = \int x^k \, dF_X(x)$$

- $m_1 = EX$ and $m_2 = EX^2$ and so on
- We define the k^{th} central moment of X by

$$s_k = E[(X - EX)^k] = \int (x - EX)^k \, dF_X(x)$$

- $s_1 = 0$ and $s_2 = \text{Var}(X)$.
- We say moments exist only when they are finite.
- Not all moments may exist for a given random variable.

- **Theorem**: If $E\left[|X|^k\right] < \infty$ then $E\left[|X|^s\right] < \infty$ for $0 < s < k$.
- For example, if third order moment exists then so do first and second order moments

Moment generating function

- The moment generating function (mgf) of rv X, $M_X : \Re \rightarrow \Re$, is defined by

$$M_X(t) = Ee^{tX} = \sum_i e^{tx_i} f_X(x_i) \text{ or } \int e^{tx} f_X(x)\, dx, \quad t \in \Re$$

- We say the mgf exists if $E[e^{tX}] < \infty$ for t in some interval around zero
- The mgf may not exist for some random variables.

- The mgf of X is: $M_X(t) = E[e^{tX}]$.
- If $M_X(t)$ exists (for $t \in [-a,\ a]$ for some $a > 0$) then all its derivatives also exist.
- Then we can get the moments of X by successive differentiation of $M_X(t)$.

$$\left.\frac{dM_X(t)}{dt}\right|_{t=0} = \left.\frac{d}{dt} E\left[e^{tX}\right]\right|_{t=0} = E[Xe^{tX}]\big|_{t=0} = EX$$

- In general

$$\left.\frac{d^k M_X(t)}{dt^k}\right|_{t=0} = E[X^k]$$

- We can easily see this by expanding e^{tX} in Taylor series:

$$\begin{aligned} M_X(t) &= Ee^{tX} = E\left[1 + \frac{tX}{1!} + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \frac{t^4X^4}{4!} + \cdots\right] \\ &= 1 + \frac{t}{1!}EX + \frac{t^2}{2!}EX^2 + \frac{t^3}{3!}EX^3 + \frac{t^4}{4!}EX^4 + \cdots \end{aligned}$$

- Now we can do term-wise differentiation. For example

$$\frac{d^3M_X(t)}{dt^3} = 0+0+0+\frac{3*2*1*t^0}{3!}EX^3+\frac{4*3*2*t}{4!}EX^4+\cdots$$

- Hence we get

$$\left.\frac{d^3M_X(t)}{dt^3}\right|_{t=0} = E[X^3]$$

Example – Moment generating function for Poisson

- $f_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \; k = 0, 1, \cdots$

$$\begin{aligned} M_X(t) &= E[e^{tX}] = \sum_{k=0}^{\infty} e^{tk} \, \frac{\lambda^k}{k!} \, e^{-\lambda} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{1}{k!} \left(\lambda e^t\right)^k \\ &= e^{-\lambda} \, e^{\lambda e^t} = e^{\lambda(e^t - 1)} \end{aligned}$$

- Now, by differentiating it we can find EX

$$EX = \left. \frac{dM_X(t)}{dt} \right|_{t=0} = e^{\lambda(e^t-1)} \, \lambda e^t \Big|_{t=0} = \lambda$$

- For mgf to exist we need $E[e^{tX}] < \infty$ for $t \in [-a,\ a]$ for some $a > 0$.
- If $M_X(t)$ exists then all moments of X are finite.
- However, all moments may be finite but the mgf may not exist.
- When mgf exists, it uniquely determines the df
- We are not saying moments uniquely determine the distribution; we are saying mgf uniquely determines the distribution

quantiles

- Let $p \in (0,\ 1)$. The number $x \in \Re$ that satisfies

$$P[X \leq x] \geq p \quad \text{and} \quad P[X \geq x] \geq 1 - p$$

  is called the quantile of order p or the $100p^{th}$ percentile of rv X.
- Suppose x is a quantile of order p. Then we have

$$p \leq F_X(x) \leq p + P[X = x]$$

- Note that for a given p there can be multiple values for x to satisfy the above.

Median of a distribution

- For $p = 0.5$, quantile of order p is called the median.
- For a continuous rv, median, x satisfies: $F_X(x) = 0.5$.
- For a discrete rv, it satisfies:
  $0.5 \leq F_X(x) \leq 0.5 + P[X = x]$.
- Median need not be unique.

- If we want to find c to minimize $E\left[(X-c)^2\right]$ then the solution is $c = EX$.
- We saw this earlier.
- Suppose we want to find c to minimize $E\left[|(X-c)|\right]$
- Then we would get c to be the median.

Mode of a Distribution

- The value of x where $f_X(x)$ attains its maximum value is called the mode of a distribution.
- For a discrete random variable it is the value that the random variable takes with highest probability.
- Take X to be binomial (with parameters, n, p). Then the mode gives the 'most probable' number of heads when we toss this coin n times.
- For a continuous rv, we can say mode is the point with 'maximum likelihood'
- In general, the mode may not be unique.
- For the Gaussian density, the mode, the median and the mean are all same.

- We next consider some inequalities involving moments of a random variable.
- These help us bound the probabilities of some important events in terms of the moments.

Markov and Chebyshev Inequalities

- Markov Inequality:

$$P[|X| > c] \leq \frac{E\left[|X|^k\right]}{c^k}$$

- Take $|X|$ as $|X - EX|$ and take $k = 2$

$$P[|X - EX| > c] \leq \frac{E\left[|X - EX|^2\right]}{c^2} = \frac{\text{Var}(X)}{c^2}$$

- This is known as the Chebyshev inequality.
- An example of what are called concentration inequalities.

- The Chebyshev inequality is

$$P[|X - EX| > c] \leq \frac{\text{Var}(X)}{c^2}$$

- Let $EX = \mu$ and let $\text{Var}(X) = \sigma^2$. Take $c = k\sigma$ (We call, σ, square root of variance, as standard deviation).
- Now, Chebyshev inequality gives us

$$P[|X - \mu| > k\sigma] \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

- This is true for all random variables and the RHS above does not depend on the distribution of X.

Proof of Markov Inequality

- Let $g : \Re \rightarrow \Re$ be a non-negative function. Then
$$P[g(X) > c] \leq \frac{E[g(X)]}{c}, \quad (c > 0)$$
(Under the assumption that the expectation is finite)
- **Proof**: We prove it for continuous rv. Proof is similar for discrete rv

$$\begin{aligned}
E[g(X)] &= \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx \\
&= \int_{g(x) \leq c} g(x)\, f_X(x)\, dx + \int_{g(x) > c} g(x)\, f_X(x)\, dx \\
&\geq \int_{g(x) > c} g(x)\, f_X(x)\, dx \quad \text{because } g(x) \geq 0 \\
&\geq c \int_{g(x) > c} f_X(x)\, dx = c\, P[g(X) > c]
\end{aligned}$$

Thus, $P[g(X) > c] \leq \frac{E[g(X)]}{c}$

Proof of Markov Inequality

$$P[g(X) > c] \leq \frac{E[g(X)]}{c}, \quad (c > 0)$$

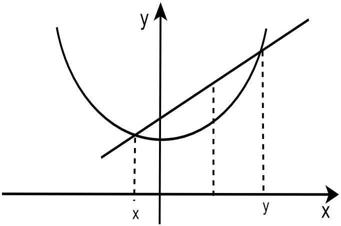
- Let $g(x) = |x|^k$ where k is a positive integer. We have $g(x) \geq 0, \ \forall x$. Let $c > 0$.
- We know that $|x| > c \Rightarrow |x|^k > c^k$ and vice versa.
- Now we get,

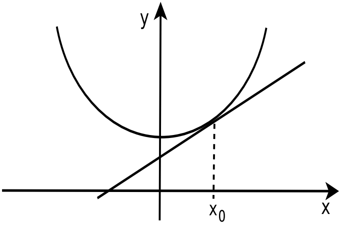
$$P[|X| > c] = P[|X|^k > c^k] \leq \frac{E\left[|X|^k\right]}{c^k}$$

(For what k is this true?)

Jensen's Inequality

- Let $g : \Re \rightarrow \Re$ be a convex function. Then
$$g(EX) \leq E[g(X)]$$
- For example, $(EX)^2 \leq E\left[X^2\right]$
- Function g is convex if (see figure on left)
$$g(\alpha x+(1-\alpha)y) \leq \alpha g(x)+(1-\alpha)g(y), \ \ \forall x, y, \ \ \forall 0 \leq \alpha \leq 1$$
- If g is convex, then, given any x_0, exists $\lambda(x_0)$ such that (see figure on right)
$$g(x) \geq g(x_0) + \lambda(x_0)(x - x_0), \ \forall x$$





Jensen's Inequality: Proof

- We have: $\forall x_0, \;\; \exists \lambda(x_0)$ such that

$$g(x) \geq g(x_0) + \lambda(x_0)(x - x_0), \; \forall x$$

- Take $x_0 = EX$ and $x = X(\omega)$. Then

$$g(X(\omega)) \geq g(EX) + \lambda(EX)(X(\omega) - EX), \; \forall \omega$$

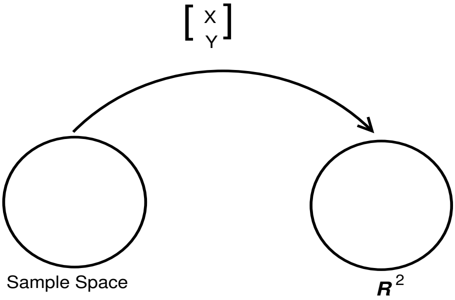
- $Y(\omega) \geq Z(\omega), \; \forall \omega \;\; \Rightarrow \;\; Y \geq Z \;\; \Rightarrow EY \geq EZ$
  Hence we get

$$\begin{aligned} g(X) \;\; &\geq \;\; g(EX) + \lambda(EX)(X - EX) \\ \Rightarrow \;\; E[g(X)] \;\; &\geq \;\; g(EX) + \lambda(EX)\; E[X - EX] = g(EX) \end{aligned}$$

- This completes the proof

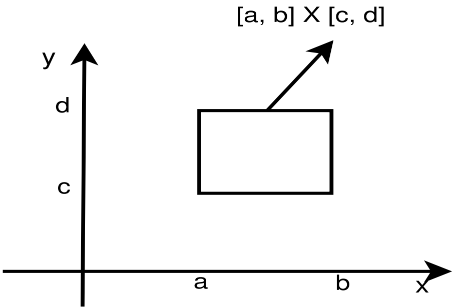
A pair of random variables

- Let X, Y be random variables on the same probability space $(\Omega, \mathcal{F}, P)$
- Each of X, Y maps Ω to $\Re$.
- We can think of the pair of radom variables as a vector-valued function that maps Ω to $\Re^2$.



- Just as in the case of a single rv, we can think of the induced probability space for the case of a pair of rv's too.
- The new sample space is $\Re^2$.
- The events now would be subsets of $\Re^2$.
  They will be borel subsets of $\Re^2$.

- Recall that Borel sets of $\Re$ are intervals and all sets that can built from intervals using countable set operations..
- Let $I_1, I_2 \subset \Re$ be intervals. Then $I_1 \times I_2 \subset \Re^2$ is known as a cylindrical set.



- Cylindrical sets are the analogues of intervals in $\Re^2$.
- Borel sets of $\Re^2$ contains all such cylindrical sets and all others that can be built using countable set operations.

Joint distribution of a pair of random variables

- Let X, Y be random variables on the same probability space $(\Omega, \mathcal{F}, P)$
- The joint distribution function of X, Y is $F_{XY} : \Re^2 \to \Re$, defined by

$$\begin{aligned} F_{XY}(x, y) &= P[X \leq x, Y \leq y] \\ &= P\left(\{\omega \ : \ X(\omega) \leq x\} \cap \{\omega \ : \ Y(\omega) \leq y\}\right) \end{aligned}$$

- The joint distribution function is the probability of the intersection of the events $[X \leq x]$ and $[Y \leq y]$.

- Recall that, for the case of a single rv, given $x_1 < x_2$, we have
$$P[x_1 < X \leq x_2] = F_X(x_2) - F_X(x_1)$$
- As we said the analogues of intervals in $\Re^2$ are cylindrical sets.
We can show, for $x_1 < x_2$, $y_1 < y_2$,

$$\begin{aligned} P[x_1 < X \leq x_2, y_1 < Y \leq y_2] &= F_{XY}(x_2, y_2) - F_{XY}(x_2, y_1) \\ &\quad - F_{XY}(x_1, y_2) + F_{XY}(x_1, y_1) \end{aligned}$$

Properties of Joint Distribution Function

- Joint distribution function: $F_{XY} : \Re^2 \rightarrow \Re$

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

- It satisfies
  1. $F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0, \forall x, y;$
     $F_{XY}(\infty, \infty) = 1$
  2. F_{XY} is non-decreasing in each of its arguments
  3. F_{XY} is right continuous and has left-hand limits in each of its arguments
  4. For all $x_1 < x_2$ and $y_1 < y_2$

$$F_{XY}(x_2, y_2) - F_{XY}(x_2, y_1) - F_{XY}(x_1, y_2) + F_{XY}(x_1, y_1) \geq 0$$

- Any $F : \Re^2 \rightarrow \Re$ satisfying the above would be a joint distribution function.

- Let X, Y be two discrete random variables (defined on the same probability space).
- Let $X \in \{x_1, \cdots x_n\}$ and $Y \in \{y_1, \cdots, y_m\}$.
- We define the joint probability mass function of X and Y as
$$f_{XY}(x_i, y_j) = P[X = x_i, Y = y_j]$$
($f_{XY}(x, y)$ is zero for all other values of x, y)
- The f_{XY} would satisfy
  - $f_{XY}(x, y) \geq 0, \ \forall x, y$ and $\sum_i \sum_j f_{XY}(x_i, y_j) = 1$
- This is a straight-forward extension of the pmf of a single discrete rv.

Example

- Consider the random experiment of rolling two dice.
  $\Omega = \{(\omega_1, \omega_2) \ : \ \omega_1, \omega_2 \in \{1, 2, \cdots, 6\}\}$
- Let X be the maximum of the two numbers and let Y be the sum of the two numbers.
  That is, $\quad X : \Omega \to \Re$, and $Y : \Omega \to \Re$ with

$$X(\omega_1, \omega_2) = \max(\omega_1, \omega_2), \quad Y(\omega_1, \omega_2) = \omega_1 + \omega_2$$

- Easy to see $X \in \{1, 2, \cdots, 6\}$ and $Y \in \{2, 3, \cdots, 12\}$

Example

- $\Omega = \{(\omega_1, \omega_2) \,:\, \omega_1, \omega_2 \in \{1, 2, \cdots, 6\}\}$
- $X(\omega_1, \omega_2) = \max(\omega_1, \omega_2), \quad Y(\omega_1, \omega_2) = \omega_1 + \omega_2$
- $X \in \{1, 2, \cdots, 6\}$ and $Y \in \{2, 3, \cdots, 12\}$
- What is the event $[X = m, Y = n]$? (We assume m, n are in the correct range)

$$[X = m, Y = n] = \{(\omega_1, \omega_2) \in \Omega \,:\, \max(\omega_1, \omega_2) = m, \ \omega_1 + \omega_2 = n\}$$

- For this to be a non-empty set, we must have $m < n \leq 2m$
- Then $[X = m, Y = n] = \{(m, \ n - m), \ (n - m, \ m)\}$
- Is this always true? No! What if $n = 2m$?
  $[X = 3, Y = 6] = \{(3, \ 3)\}$,
  $[X = 4, Y = 6] = \{(4, \ 2), \ (2, \ 4)\}$
- So, $P[X = m, Y = n]$ is either $2/36$ or $1/36$ (assuming m, n satisfy other requirements)

Example

- We can now write the joint pmf.
- Assume $1 \leq m \leq 6$ and $2 \leq n \leq 12$. Then

$$f_{XY}(m,n) = \begin{cases} \frac{2}{36} & \text{if } \; m < n < 2m \\ \frac{1}{36} & \text{if } \; n = 2m \end{cases}$$

($f_{XY}(m,n)$ is zero in all other cases)

- Does this satisfy requirements of joint pmf?

$$\begin{aligned} \sum_{m,n} f_{XY}(m,n) &= \sum_{m=1}^{6} \sum_{n=m+1}^{2m-1} \frac{2}{36} + \sum_{m=1}^{6} \frac{1}{36} \\ &= \frac{2}{36} \sum_{m=1}^{6} (m-1) + \frac{1}{36} 6 \\ &= \frac{2}{36}(21-6) + \frac{6}{36} = 1 \end{aligned}$$

Joint Probability mass function

- Let $X \in \{x_1, x_2, \cdots\}$ and $Y \in \{y_1, y_2, \cdots\}$ be discrete random variables.
- The joint pmf: $f_{XY}(x, y) = P[X = x, Y = y]$.
- The joint pmf satisfies:
  - $f_{XY}(x, y) \geq 0, \forall x, y$ and
  - $\sum_i \sum_j f_{XY}(x_i, y_j) = 1$
- Given the joint pmf, we can get the joint df as

$$F_{XY}(x, y) = \sum_{\substack{i: \\ x_i \leq x}} \sum_{\substack{j: \\ y_j \leq y}} f_{XY}(x_i, y_j)$$

- We normally specify joint pmf

- Given the joint pmf, we can (in principle) compute the probability of any event involving the two discrete random variables.

$$P[(X, Y) \in B] = \sum_{\substack{i,j: \\ (x_i, y_j) \in B}} f_{XY}(x_i, y_j)$$

- Now, events can be specified in terms of relations between the two rv's too.
  For example,

$$[X < Y + 2] = \{\omega \ : \ X(\omega) < Y(\omega) + 2\}$$

- Thus,

$$P[X < Y + 2] = \sum_{\substack{i,j: \\ x_i < y_j + 2}} f_{XY}(x_i, y_j)$$

- Take the example: 2 dice, X is max and Y is sum
- $f_{XY}(m,n) = 0$ unless $m = 1, \cdots, 6$ and $n = 2, \cdots, 12$. For this range

$$f_{XY}(m,n) = \begin{cases} \frac{2}{36} & \text{if } \; m < n < 2m \\ \frac{1}{36} & \text{if } \; n = 2m \end{cases}$$

- Suppose we want $P[Y = X + 2]$.

$$\begin{aligned} P[Y = X + 2] &= \sum_{\substack{m,n: \\ n = m+2}} f_{XY}(m,n) = \sum_{m=1}^{6} f_{XY}(m, m+2) \\ &= \sum_{m=2}^{6} f_{XY}(m, m+2) \quad \text{since we need } m + 2 \leq 2m \\ &= \frac{1}{36} + 4\,\frac{2}{36} = \frac{9}{36} \end{aligned}$$

Joint density function

- Let X, Y be two continuous rv's with df F_{XY}.
- If there exists a function f_{XY} that satisfies

$$F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x', y') \; dy' \; dx', \;\; \forall x, y$$

  then we say that X, Y have a joint probability density function which is f_{XY}
- Please note the difference in the definition of joint pmf and joint pdf.
- When X, Y are discrete we defined a joint pmf
- We are not saying that if X, Y are continuous rv's then a joint density exists.

properties of joint density

- The joint density (or joint pdf) of X, Y is f_{XY} that satisfies

$$F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x', y')\, dy'\, dx', \ \ \forall x, y$$

- Since F_{XY} is non-decreasing in each argument, we must have $f_{XY}(x, y) \geq 0$.
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x', y')\, dy'\, dx' = 1$ is needed to ensure $F_{XY}(\infty, \infty) = 1$.

properties of joint density

- The joint density f_{XY} satisfies the following
  1. $f_{XY}(x, y) \geq 0, \ \forall x, y$
  2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x', y') \, dy' \, dx' = 1$
- Ay function that satisfies these two is a joint density.
- These are very similar to the properties of the density of a single rv

Example: Joint Density

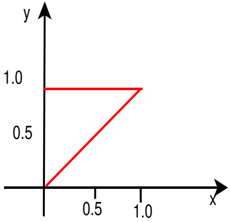
- Consider the function

$$f(x, y) = 2, \; 0 < x < y < 1 \;\; (f(x, y) = 0, \; \text{otherwise})$$

- Let us show this is a density

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \; dx \; dy = \int_0^1 \int_0^y 2 \; dx \; dy = \int_0^1 2 \; x|_0^y \; dy = \int_0^1 2y \; dy = 1$$

- We can say this density is uniform over the region



The figure is not a plot of the density function!!

- Joint density function is $f_{XY} : \Re^2 \to \Re$ that satisfies
  - $f_{XY}(x, y) \geq 0, \ \forall x, y$
  - $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) \ dx \ dy = 1$
- The Joint distribution function is given by $F_{XY}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(x, y) \ dx \ dy$
- We specify a pair of continuous rv with joint density (when it exists)
- We also have

$$P[x_1 \leq X \leq x_2, \ y_1 \leq Y \leq y_2] = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{XY} \ dy \ dx$$

- In general

$$P[(X, Y) \in B] = \int_B f_{XY}(x, y) \ dx \ dy,$$

- Let us consider the example

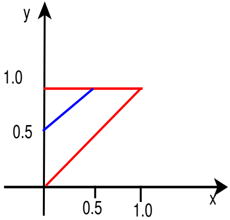
$$f(x,y) = 2, \;\; 0 < x < y < 1$$

- Suppose wee want probability of $[Y > X + 0.5]$

$$\begin{aligned} P[Y > X + 0.5] &= P\left[(X, Y) \in \{(x,y) : y > x + 0.5\}\right] \\ &= \int_{\{(x,y)\,:\,y > x+0.5\}} f_{XY}(x,y)\; dx\; dy \\ &= \int_{0.5}^{1} \int_{0}^{y-0.5} 2 \; dx\; dy \\ &= \int_{0.5}^{1} 2(y - 0.5) dy \end{aligned}$$

$$= 2\left.\frac{y^2}{2}\right|_{0.5}^{1} - y|_{0.5}^{1} = 1 - 0.25 - 1 + 0.5 = 0.25$$

- We can look at it geometrically



- The probability of the event we want is the area of the small triangle divided by that of the big triangle.

Marginal Distributions

- Let X, Y be random variables with joint distribution function F_{XY}.
- We know $F_{XY}(x, y) = P[X \leq x, Y \leq y]$.
- Hence

$$F_{XY}(x, \infty) = P[X \leq x, Y \leq \infty] = P[X \leq x] = F_X(x)$$

- We define the marginal distribution functions of X, Y by

$$F_X(x) = F_{XY}(x, \infty); \quad F_Y(y) = F_{XY}(\infty, y)$$

- These are simply distribution functions of X and Y obtained from the joint distribution function.

Marginal mass functions

- Let $X \in \{x_1, x_2, \cdots\}$ and $Y \in \{y_1, y_2, \cdots\}$
- Let f_{XY} be their joint mass function.
- Then

$$P[X = x_i] = \sum_j P[X = x_i, Y = y_j] = \sum_j f_{XY}(x_i, y_j)$$

(This is because $[Y = y_j],\ j = 1, \cdots,$ form a partition and $P(A) = \sum_i P(AB_i)$ when B_i is a partition)

- We define the marginal mass functions of X and Y as

$$f_X(x_i) = \sum_j f_{XY}(x_i, y_j); \quad f_Y(y_j) = \sum_i f_{XY}(x_i, y_j)$$

- These are mass functions of X and Y obtained from the joint mass function

marginal density functions

- Let X, Y be continuous rv with joint density f_{XY}.
- Then we know $F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} \ f_{XY}(x', y') \ dy' \ dx'$
- Hence, we have

$$\begin{aligned} F_X(x) = F_{XY}(x, \infty) &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{XY}(x', y') \ dy' \ dx' \\ &= \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} f_{XY}(x', y') \ dy' \right) \ dx' \end{aligned}$$

- Since X is a continuous rv, this means

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) \ dy$$

  We call this the marginal density of X.
- Similarly, marginal density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) \ dx$$

- These are pdf's of X and Y obtained from the joint

Example

- Rolling two dice, X is max, Y is sum
- We had, for $1 \leq m \leq 6$ and $2 \leq n \leq 12$,

$$f_{XY}(m,n) = \begin{cases} \frac{2}{36} & \text{if} \ \ m < n < 2m \\ \frac{1}{36} & \text{if} \ \ n = 2m \end{cases}$$

- We know, $f_X(m) = \sum_n f_{XY}(m,n), \ m = 1, \cdots, 6.$
- Given m, for what values of n, $f_{XY}(m,n) > 0$?
  We can only have $n = m+1, \cdots, 2m$.
- Hence we get

$$f_X(m) = \sum_{n=m+1}^{2m} f_{XY}(m,n) = \sum_{n=m+1}^{2m-1} \frac{2}{36} + \frac{1}{36} = \frac{2}{36}(m-1) + \frac{1}{36} = \frac{2m-1}{36}$$

Example

- Consider the joint density

$$f_{XY}(x,y) = 2, \; 0 < x < y < 1$$

- The marginal density of X is: for $0 < x < 1$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\; dy = \int_x^1 2 \; dy = 2(1-x)$$

  Thus, $f_X(x) = 2(1-x), \; 0 < x < 1$
- We can easily verify this is a density

$$\int_{-\infty}^{\infty} f_X(x)\; dx = \int_0^1 2(1-x)\; dx = (2x - x^2)\Big|_0^1 = 1$$

We have: $f_{XY}(x, y) = 2, \;\; 0 < x < y < 1$

- We can similarly find density of Y.
- For $0 < y < 1$,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\; dx = \int_0^y 2\; dx = 2y$$

- Thus, $f_Y(y) = 2y, \; 0 < y < 1$ and

$$\int_0^1 2y\; dy = 2\, \frac{y^2}{2}\bigg|_0^1 = 1$$

- If we are given the joint df or joint pmf/joint density of X, Y, then the individual df or pmf/pdf are uniquely determined.
- However, given individual pdf of X and Y, we cannot determine the joint density. (same is true of pmf or df)
- There can be many different joint density functions all having the same marginals
- With the same values for $P(A), P(B)$, there can be many different values for $P(AB)$.

Conditional distributions

- Let X, Y be rv's on the same probability space
- We define the conditional distribution function of X given Y by
$$F_{X|Y}(x|y) = P[X \leq x | Y = y]$$
This is well defined whenever $f_Y(y) > 0$.
- Note that $F_{X|Y} : \Re^2 \to \Re$
- $F_{X|Y}(x|y)$ is a notation. We could write $F_{X|Y}(x, y)$.

- Conditional distribution function of X given Y is

$$F_{X|Y}(x|y) = P[X \leq x | Y = y]$$

It is the conditional probability of $[X \leq x]$ given (or conditioned on) $[Y = y]$.

- Consider example: rolling 2 dice, X is max, Y is sum

$$P[X \leq 4 | Y = 3] = 1; \;\; P[X \leq 4 | Y = 9] = 0$$

- This is what conditional distribution captures.
- For every value of y, $F_{X|Y}(x|y)$ is a distribution function in the variable x.
- It defines a new distribution for X based on knowing the value of Y.

Conditional mass function

- We define the conditional mass function of X given Y as

$$f_{X|Y}(x_i|y_j) = \frac{f_{XY}(x_i, y_j)}{f_Y(y_j)} = P[X = x_i | Y = y_j]$$

- Note that

$$\sum_i f_{X|Y}(x_i|y_j) = 1, \ \forall y_j; \quad \text{and} \quad F_{X|Y}(x|y_j) = \sum_{i:x_i \leq x} f_{X|Y}(x_i|y_j)$$

Example: Conditional pmf

- Consider the random experiment of tossing a coin n times.
- Let X denote the number of heads and let Y denote the toss number on which the first head comes.
- For $1 \leq k \leq n$

$$\begin{aligned} f_{Y|X}(k|1) &= P[Y = k|X = 1] = \frac{P[Y = k, X = 1]}{P[X = 1]} \\ &= \frac{p(1-p)^{n-1}}{{}^nC_1 \; p(1-p)^{n-1}} \\ &= \frac{1}{n} \end{aligned}$$

- Given there is only one head, it is equally likely to occur on any toss.

- The conditional mass function is

$$f_{X|Y}(x_i|y_j) = P[X = x_i|Y = y_j] = \frac{f_{XY}(x_i, y_j)}{f_Y(y_j)}$$

- This gives us the useful identity

$$f_{XY}(x_i, y_j) = f_{X|Y}(x_i|y_j) f_Y(y_j)$$

$$(\ P[X = x_i, Y = y_j] = P[X = x_i|Y = y_j] P[Y = y_j])$$

- This gives us the total proability rule for discrete rv's

$$f_X(x_i) = \sum_j f_{XY}(x_i, y_j) = \sum_j f_{X|Y}(x_i|y_j) f_Y(y_j)$$

- This is same as

$$P[X = x_i] = \sum_j P[X = x_i|Y = y_j] P[Y = y_j]$$

$(P(A) = \sum_j P(A|B_j)P(B_j)$ when $B_1, \cdots$ form a partition)

Bayes Rule for discrete Random Variable

- We have

$$f_{XY}(x_i, y_j) = f_{X|Y}(x_i|y_j)f_Y(y_j) = f_{Y|X}(y_j|x_i)f_X(x_i)$$

- This gives us Bayes rule for discrete rv's

$$\begin{aligned} f_{X|Y}(x_i|y_j) &= \frac{f_{Y|X}(y_j|x_i)f_X(x_i)}{f_Y(y_j)} \\ &= \frac{f_{Y|X}(y_j|x_i)f_X(x_i)}{\sum_i f_{XY}(x_i, y_j)} \\ &= \frac{f_{Y|X}(y_j|x_i)f_X(x_i)}{\sum_i f_{Y|X}(y_j|x_i)f_X(x_i)} \end{aligned}$$

- Let X, Y be continuous rv's with joint density, f_{XY}.
- We once again want to define conditional df

$$F_{X|Y}(x|y) = P[X \leq x|Y = y]$$

- But the conditioning event, $[Y = y]$ has zero probability.
- Hence we define conditional df as follows

$$F_{X|Y}(x|y) \;\; = \;\; \lim_{\delta \downarrow 0} P[\, X \leq x \mid Y \in [y, \; y + \delta] \,]$$

- This is well defined if the limit exists.
- The limit exists for all y where $f_Y(y) > 0$ (and for all x)

- The conditional df is given by (assuming $f_Y(y) > 0$)
$F_{X|Y}(x|y) = \lim_{\delta \downarrow 0} P[X \leq x \mid Y \in [y,\ y + \delta]\,]$
- By calculating the limit, we can show that

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} \frac{f_{XY}(x', y)}{f_Y(y)}\, dx'$$

- We define conditional density of X given Y as

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

- Similarly, conditional density of Y given X is

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

Example

$$f_{XY}(x,y) = 2,\ 0 < x < y < 1$$

- We saw that the marginal densities are

$$f_X(x) = 2(1-x),\ 0 < x < 1; \quad f_Y(y) = 2y,\ 0 < y < 1$$

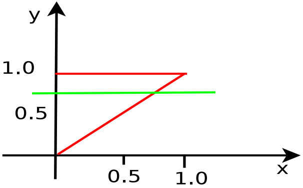
- Hence the conditional densities are given by

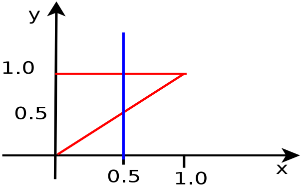
$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{1}{y},\ 0 < x < y < 1 \\ f_{Y|X}(y|x) &= \frac{f_{XY}(x,y)}{f_X(x)} = \frac{1}{1-x},\ 0 < x < y < 1 \end{aligned}$$

- $f_{XY}(x, y) = 2, \ 0 < x < y < 1$

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{1}{y}, \ 0 < x < y < 1 \\ f_{Y|X}(y|x) &= \frac{1}{1-x}, \ 0 < x < y < 1 \end{aligned}$$

- We can see this intuitively
  Conditioned on $Y = y$, X is uniform over $(0, y)$.
  Conditioned on $X = x$, Y is uniform over $(x, 1)$.





- Let X, Y have joint density f_{XY}.
- The conditional df of X given Y is

$$F_{X|Y}(x|y) = \lim_{\delta \downarrow 0} P[X \leq x | Y \in [y, \ y + \delta]]$$

- This exists if $f_Y(y) > 0$ and then it has a density:

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} f_{X|Y}(x'|y) \ dx'$$

- This conditional density is given by

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

- We (once again) have the useful identity

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

- The identity $f_{XY}(x, y) = f_{X|Y}(x|y)f_Y(y)$ can be used to specify the joint density of two continuous rv's
- We can specify the marginal density of one and the conditional density of the other given the first.
- This may actually be the model of how the the rv's are generated.

Example

- Let X be uniform over $(0,\ 1)$ and let Y be uniform over 0 to X. Find the density of Y.
- What we are given is

$$f_X(x) = 1,\ 0 < x < 1; \quad f_{Y|X}(y|x) = \frac{1}{x}, 0 < y < x < 1$$

- Hence the joint density is: $f_{XY}(x, y) = \frac{1}{x},\ 0 < y < x < 1.$
- Hence the density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\ dx = \int_y^1 \frac{1}{x}\ dx = -\ln(y),\ 0 < y < 1$$

- We can verify it to be a density

$$-\int_0^1 \ln(y)\ dy = -y\ln(y)|_0^1 + \int_0^1 y\,\frac{1}{y}\ dy = 1$$

- We have the identity

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

- By integrating both sides

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dy = \int_{-\infty}^{\infty} f_{X|Y}(x|y)\, f_Y(y)\, dy$$

- This is a continuous analogue of total probability rule.
- But note that, since X is continuous rv, $f_X(x)$ is **NOT** $P[X = x]$
- In case of discrete rv, we had

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

- It is as if one can simply replace pmf by pdf and summation by integration!!
- While often that gives the right result, one needs to be very careful

- We have the identity

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

- This gives rise to Bayes rule for continuous rv

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} \\ &= \frac{f_{Y|X}(y|x) f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x)\ dx} \end{aligned}$$

- This is essentially identical to Bayes rule for discrete rv's. We have essentially put the pdf wherever there was pmf

- To recap, we started by defining conditional distribution function.

$$F_{X|Y}(x|y) = P[X \leq x \mid Y = y]$$

- When X, Y are discrete, this is well defined when $f_Y(y) > 0$.
  That is, we define it only for all values that Y can take.
- When X, Y have joint density, we defined it by

$$F_{X|Y}(x|y) = \lim_{\delta \downarrow 0} P[X \leq x \mid Y \in [y, y + \delta]\,]$$

  This limit exists and $F_{X|Y}$ is well defined if $f_Y(y) > 0$.
  That is, essentially again for all values that Y can take.
- In the discrete case, we define $f_{X|Y}$ as the pmf corresponding to $F_{X|Y}$.
- In the continuous case $f_{X|Y}$ is the density corresponding to $F_{X|Y}$.
- In both cases we have: $f_{XY}(x, y) = f_{X|Y}(x|y)f_Y(y)$
- This gives total probability rule and Bayes rule for random variables

- Now, let X be a continuous rv and let Y be discrete rv. Now we cannot have joint pmf or pdf.
- But, we can still define $F_{X|Y}$ as

$$F_{X|Y}(x|y) = P[X \leq x | Y = y]$$

This is well defined when $f_Y(y) > 0$.

- Since X is continuous rv, this df would have a density

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} f_{X|Y}(x'|y) \, dx'$$

- Hence we can write

$$\begin{aligned} P[X \leq x, Y = y] &= F_{X|Y}(x|y) P[Y = y] \\ &= \int_{-\infty}^{x} f_{X|Y}(x'|y) \, f_Y(y) \, dx' \end{aligned}$$

- We now get

$$\begin{aligned} F_X(x) &= P[X \leq x] = \sum_y P[X \leq x, Y = y] \\ &= \sum_y \int_{-\infty}^{x} f_{X|Y}(x'|y)\ f_Y(y)\ dx' \\ &= \int_{-\infty}^{x} \sum_y f_{X|Y}(x'|y)\ f_Y(y)\ dx' \end{aligned}$$

- This gives us

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

- This is another version of total probability rule.
- Earlier we derived this when X, Y are discrete.
- The formula is true even when X is continuous
  Only difference is we need to take f_X as the density of X.

- When X is continuous and Y is discrete, we defined $f_{X|Y}(x|y)$ to be the density corresponding to $F_{X|Y}(x|y) = P[X \leq x | Y = y]$
- Then we once again get

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

  Here f_X is density (and not a mass function). $f_{X|Y}$ is also a density.
- This is an interesting way to specify a density.
- Suppose $Y \in \{1, 2, 3\}$.
  Let $f_Y(i) = \lambda_i$. $(\lambda_i \geq 0, \; \sum_i \lambda_i = 1)$
  Let $f_{X|Y}(x|i) = f_i(x)$. Then

$$f_X(x) = \lambda_1 f_1(x) + \lambda_2 f_2(x) + \lambda_3 f_3(x)$$

  Called a mixture density model

- Continuing with X continuous rv and Y discrete
- Can we define $f_{Y|X}(y|x)$? $\;(P[Y = y \mid X = x]\ ?)$
- We can define it as

$$f_{Y|X}(y|x) = \lim_{\delta \downarrow 0} P[Y = y \mid X \in [x, x + \delta]\,]$$

- By simplifying this we get the following:

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

- This gives us

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

- Since $\int_{-\infty}^{\infty} f_{X|Y}(x|y)\ dx = 1$, we get

$$\int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x) = f_Y(y)$$

Another version of total probability rule.

- Earlier we saw total probability rule and Bayes rule versions when either X, Y have a joint pmf or have a joint pdf.
- We can now extend them to the case when one of them is continuous rv and the other is discrete rv.

- Let us review all the total probability formulas

$$\textbf{1. } f_X(x) = \sum_y f_{X|Y}(x|y) f_Y(y)$$

- We first derived this when X, Y are discrete.
- But this holds whenever Y is discrete
If X is continuous the $f_X, f_{X|Y}$ are densities; If X is also discrete they are mass functions

$$\textbf{2. } f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x)\, dx$$

- We first proved it when X, Y have a joint density
This also holds when X is cont and Y is discrete. In that case f_Y is a mass function

- When X is continuous rv and Y is discrete rv, we derived

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

- This once again gives rise to Bayes rule:

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

- Earlier we showed this hold when X, Y are both discrete or both continuous.
- Thus Bayes rule holds in all four possible scenarios
- Only difference is we need to interpret f_X or $f_{X|Y}$ as mass functions when X is discrete and as densities when X is a continuous rv
- In general, one refers to these always as densities since the actual meaning would be clear from context.

Example

- Consider a communication system. The transmitter puts out 0 or 5 volts for the bits of 0 and 1, and, volage measured by the receiver is the sent voltage plus noise added by the channel.
- We assume noise has Gaussian density with mean zero and variance σ^2.
- We want the probability that the sent bit is 1 when measured voltage at the receiver is x. (This is for deciding what is sent).
- Let X be the measured voltage and let Y be sent bit.
- We want to calculate $f_{Y|X}(1|x)$.
- We want to use the Bayes rule to calculate this

- We need $f_{X|Y}$. What does our model say?
- $f_{X|Y}(x|1)$ is Gaussian with mean 5 and variance σ^2 and $f_{X|Y}(x|0)$ is Gaussian with mean zero and variance σ^2

$$P[Y = 1|X = x] = f_{Y|X}(1|x) = \frac{f_{X|Y}(x|1)\ f_Y(1)}{f_X(x)}$$

- We need $f_Y(1), f_Y(0)$. Let us take them to be same.
- In practice we only want to know whether $f_{Y|X}(1|x) > f_{Y|X}(0|x)$
- Then we do not need to calculate $f_X(x)$.
  We only need ratio of $f_{Y|X}(1|x)$ and $f_{Y|X}(0|x)$.

- The ratio of the two probabilities is

$$
\begin{aligned}
\frac{f_{Y|X}(1|x)}{f_{Y|X}(0|x)} &= \frac{f_{X|Y}(x|1)\ f_Y(1)}{f_{X|Y}(x|0)\ f_Y(0)} \\
&= \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-5)^2}}{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-0)^2}} \\
&= e^{-0.5\sigma^{-2}(x^2-10x+25-x^2)} \\
&= e^{0.5\sigma^{-2}(10x-25)}
\end{aligned}
$$

- We are only interested in whether the above is greater than 1 or not.
- The ratio is greater than 1 if $10x > 25$ or $x > 2.5$
- So, if $X > 2.5$ we will conclude bit 1 is sent. Intuitively obvious!

- We did not calculate $f_X(x)$ in the above.
- We can calculate it if we want.
- Using total probability rule

$$
\begin{aligned}
f_X(x) &= \sum_y f_{X|Y}(x|y) f_Y(y) \\
&= f_{X|Y}(x|1) f_Y(1) + f_{X|Y}(x|0) f_Y(0) \\
&= \frac{1}{2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-5)^2}{2\sigma^2}} + \frac{1}{2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}
\end{aligned}
$$

- It is a mixture density

Independent Random Variables

- Two random variable X, Y are said to be independent if for all B_1, B_2, the events $[X \in B_1]$ and $[Y \in B_2]$ are independent.
- If X, Y are independent then

$$P[X \in B_1, Y \in B_2] = P[X \in B_1] \; P[Y \in B_2], \;\; \forall B_1, B_2$$

- In particular

$$F_{XY}(x, y) = P[X \leq x, Y \leq y] = P[X \leq x] P[Y \leq y] = F_X(x) \; F_Y(y)$$

- **Theorem**: X, Y are independent if and only if $F_{XY}(x, y) = F_X(x) F_Y(y)$.
  We will not prove this theorem here.

- Suppose X, Y are independent discrete rv's

$$f_{XY}(x,y) = P[X = x, Y = y] = P[X = x]P[Y = y] = f_X(x)f_Y(y)$$

  The joint mass function is a product of marginals.

- Suppose $f_{XY}(x,y) = f_X(x)f_Y(y)$. Then

$$\begin{aligned} F_{XY}(x,y) &= \sum_{x_i \leq x, y_j \leq y} f_{XY}(x_i, y_j) = \sum_{x_i \leq x, y_j \leq y} f_X(x_i)f_Y(y_j) \\ &= \sum_{x_i \leq x} f_X(x_i) \sum_{y_j \leq y} f_Y(y_j) = F_X(x)F_Y(y) \end{aligned}$$

- So, X, Y are independent if and only if $f_{XY}(x,y) = f_X(x)f_Y(y)$

- Let X, Y be independent continuous rv

$$\begin{aligned} F_{XY}(x,y) &= F_X(x)F_Y(y) = \int_{-\infty}^{x} f_X(x')\,dx' \int_{-\infty}^{y} f_Y(y')\,dy' \\ &= \int_{-\infty}^{y}\int_{-\infty}^{x} (f_X(x')f_Y(y'))\; dx'\; dy' \end{aligned}$$

- This implies joint density is product of marginals.
- Now, suppose $f_{XY}(x,y) = f_X(x)f_Y(y)$

$$\begin{aligned} F_{XY}(x,y) &= \int_{-\infty}^{y}\int_{-\infty}^{x} f_{XY}(x',y')\,dx'\,dy' \\ &= \int_{-\infty}^{y}\int_{-\infty}^{x} f_X(x')f_Y(y')\,dx'\,dy' \\ &= \int_{-\infty}^{x} f_X(x')\,dx' \int_{-\infty}^{y} f_Y(y')\,dy' = F_X(x)F_Y(y) \end{aligned}$$

- So, X, Y are independent if and only if $f_{XY}(x,y) = f_X(x)f_Y(y)$

- Let X, Y be independent.
- Then $P[X \in B_1 | Y \in B_2] = P[X \in B_1]$.
- Hence, we get $F_{X|Y}(x|y) = F_X(x)$.
- This also implies $f_{X|Y}(x|y) = f_X(x)$.
- This is true for all the four possibilities of X, Y being continuous/discrete.

More than two rv

- Everything we have done so far is easily extended to multiple random variables.
- Let X, Y, Z be rv on the same probability space.
- We define joint distribution function by

$$F_{XYZ}(x, y, z) = P[X \leq x, Y \leq y, Z \leq z]$$

- If all three are discrete then the joint mass function is

$$f_{XYZ}(x, y, z) = P[X = x, Y = y, Z = z]$$

- If they are continuous , they have a joint density if

$$F_{XYZ}(x, y, z) = \int_{-\infty}^{z} \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XYZ}(x', y', z')\ dx'\ dy'\ dz'$$

- Easy to see that joint mass function satisfies
  1. $f_{XYZ}(x, y, z) \geq 0$ and is non-zero only for countably many tuples.
  2. $\sum_{x,y,z} \ f_{XYZ}(x, y, z) = 1$
- Similarly the joint density satisfies
  1. $f_{XYZ}(x, y, z) \geq 0$
  2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \ f_{XYZ}(x, y, z) \ dx \ dy \ dz = 1$
- These are straight-forward generalizations
- We specify multiple random variables either through joint mass function or joint density function.

- Now we get many different marginals:

$$F_{XY}(x,y) = F_{XYZ}(x,y,\infty); \;\; F_Z(z) = F_{XYZ}(\infty,\infty,z) \quad \text{and so on}$$

- Similarly we get

$$\begin{aligned} f_{YZ}(y,z) &= \int_{-\infty}^{\infty} f_{XYZ}(x,y,z)\, dx; \\ f_X(x) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XYZ}(x,y,z)\, dy\, dz \end{aligned}$$

- Any marginal is a joint density of a subset of these rv's and we obtain it by integrating the (full) joint density with respect to the remaining variables.
- We obtain the marginal mass functions for a subset of the rv's also similarly where we sum over the remaining variables.

- Like in case of marginals, there are different types of conditional distributions now.
- We can always define conditional distribution functions like

$$\begin{aligned} F_{XY|Z}(x, y|z) &= P[X \leq x, Y \leq y|Z = z] \\ F_{X|YZ}(x|y, z) &= P[X \leq x|Y = y, Z = z] \end{aligned}$$

- In all such cases, if the conditioning random variables are continuous, we define the above as a limit.
- For example when Z is continuous

$$F_{XY|Z}(x, y|z) = \lim_{\delta \downarrow 0} P[X \leq x, Y \leq y \mid Z \in [z, z + \delta]\,]$$

- If X, Y, Z are all discrete then, all conditional mass functions are defined by appropriate conditional probabilities. For example,

$$f_{X|YZ}(x|y,z) = P[X = x|Y = y, Z = z]$$

- Thus the following are obvious

$$\begin{aligned} f_{XY|Z}(x,y|z) &= \frac{f_{XYZ}(x,y,z)}{f_Z(z)} \\ f_{X|YZ}(x|y,z) &= \frac{f_{XYZ}(x,y,z)}{f_{YZ}(y,z)} \\ f_{XYZ}(x,y,z) &= f_{Z|YX}(z|y,x) f_{Y|X}(y|x) f_X(x) \end{aligned}$$

- For example, the first one above follows from

$$P[X = x, Y = y|Z = z] = \frac{P[X = x, Y = y, Z = z]}{P[Z = z]}$$

- When X, Y, Z have joint density, all such relations hold for the appropriate (conditional) densities. For example,

$$f_{XYZ}(x, y, z) = f_{Z|XY}(z|x, y) f_{XY}(x, y) = f_{Z|XY}(z|x, y) f_{Y|X}(y|x) f_X(x)$$

- We can similarly talk about the joint distribution of any finite number of rv's
- Let $X_1, X_2, \cdots, X_n$ be rv's on the same probability space.
- We denote it as a vector $\mathbf{X}$ or $\underline{X}$. We can think of it as a mapping, $\mathbf{X} : \Omega \to \Re^n$.
- We can write the joint distribution as

$$F_{\mathbf{X}}(\mathbf{x}) = P[\mathbf{X} \leq \mathbf{x}] = P[X_i \leq x_i, \; i = 1, \cdots, n]$$

- We represent by $f_{\mathbf{X}}(\mathbf{x})$ the joint density or mass function. Sometimes we also write it as $f_{X_1 \cdots X_n}(x_1, \cdots, x_n)$
- We use similar notation for marginal and conditional distributions. For example,
  $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y})$

- When some variables are continuous and others are discrete we do not have joint pmf or pdf.
- But we can always define conditional distribution functions and from there can get conditional densities (or mass functions).
- Thus total probability rule and Bayes rule hold in all such cases.

Example

- Let a joint density be given by

$$f_{XYZ}(x,y,z) = K, \quad 0 < z < y < x < 1$$

First let us determine K.

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XYZ}(x,y,z)\ dz\ dy\ dx &= \int_0^1 \int_0^x \int_0^y K\ dz\ dy\ dx \\
&= K \int_0^1 \int_0^x y\ dy\ dx \\
&= K \int_0^1 \frac{x^2}{2}\ dx \\
&= K \frac{1}{6} \quad \Rightarrow K = 6
\end{aligned}$$

$$f_{XYZ}(x, y, z) = 6, \quad 0 < z < y < x < 1$$

- Suppose we want to find the (marginal) joint distribution of X and Z.

$$\begin{aligned} f_{XZ}(x, z) &= \int_{-\infty}^{\infty} f_{XYZ}(x, y, z) \, dy \\ &= \int_{z}^{x} 6 \, dy, \quad 0 < z < x < 1 \\ &= 6(x - z), \quad 0 < z < x < 1 \end{aligned}$$

- The joint density of X, Y, Z is

$$f_{XYZ}(x, y, z) = 6, \quad 0 < z < y < x < 1$$

- The joint density of X, Z is

$$f_{XZ}(x, z) = 6(x - z), \quad 0 < z < x < 1$$

- Hence,

$$f_{Y|XZ}(y|x, z) = \frac{f_{XYZ}(x, y, z)}{f_{XZ}(x, z)} = \frac{1}{x - z}, \quad 0 < z < y < x < 1$$

Independence of multiple random variables

- Random variables $X_1, X_2, \cdots, X_n$ are said to be independent if, for any $B_1, \cdots, B_n$, the events $[X_i \in B_i],\ i = 1, \cdots, n$ are independent.
  (Recall definition of independence of a set of events)
- Independence implies that the marginals would determine the joint distribution.
- If X, Y, Z are independent then
  $f_{XYZ}(x, y, z) = f_X(x) f_Y(y) f_Z(z)$

Functions of multiple random variables

- Let X, Y be random variables on the same probability space.
- Let $g : \Re^2 \rightarrow \Re$.
- Let $Z = g(X, Y)$. Then Z is a rv
- This is analogous to functions of a single rv
- This is easily extended to function of multiple random variables.

- let $Z = g(X, Y)$
- We can determine df or pmf/pdf of Z from the joint distribution of X, Y
- For example, if X, Y are discrete, then

$$f_Z(z) = P[Z = z] = P[g(X, Y) = z] = \sum_{\substack{x_i, y_j: \\ g(x_i, y_j) = z}} f_{XY}(x_i, y_j)$$

iid random variables

- Suppose X, Y are independent and $f_X = f_Y$.
- Then they are called **independent and identically distributed** or **iid** random variables.
- Similarly for any number of random variables.
- $X_1, \cdots, X_n$ are iid means
  $X_1, \cdots, X_n$ are independent and $f_{X_i} = f$, $\forall i$.

Example

- Let $Z = \max(X, Y)$. Then we have

$$\begin{aligned} F_Z(z) &= P[Z \leq z] = P[\max(X, Y) \leq z] \\ &= P[X \leq z, Y \leq z] \\ &= F_{XY}(z, z) \\ &= F_X(z) F_Y(z), \quad \text{if } X, Y \text{ are independent} \\ &= (F_X(z))^2, \quad \text{if they are iid} \end{aligned}$$

- This is true of all iid random variables.
- Suppose X, Y are iid continuous rv. Then density of Z is

$$f_Z(z) = 2F_X(z) f_X(z)$$

- This is easily generalized to n radom variables.
- Let $Z = \max(X_1, \cdots, X_n)$

$$\begin{aligned} F_Z(z) &= P[Z \leq z] = P[\max(X_1, X_2, \cdots, X_n) \leq z] \\ &= P[X_1 \leq z, X_2 \leq z, \cdots, X_n \leq z] \\ &= F_{X_1 \cdots X_n}(z, \cdots, z) \\ &= F_{X_1}(z) \cdots F_{X_n}(z), \quad \text{if they are independent} \\ &= (F_X(z))^n, \quad \text{if they are iid} \\ & \qquad \text{where we take } F_X \text{ as the common df} \end{aligned}$$

- For example if all X_i are uniform over $(0, 1)$ and ind, then
  $F_Z(z) = z^n,\ 0 < z < 1$
  $(F_Z(z) = 0,\ z \leq 0,\ F_Z(z) = 1,\ z \geq 1)$
  $f_Z(z) = nz^{n-1},\ 0 < z < 1$

- Consider $Z = \min(X, Y)$ and X, Y independent

$$F_Z(z) = P[Z \le z] = P[\min(X, Y) \le z]$$

- It is difficult to write this in terms of joint df of X, Y.
- So, we consider the following

$$\begin{aligned} P[Z > z] &= P[\min(X, Y) > z] \\ &= P[X > z, Y > z] \\ &= P[X > z]P[Y > z], \quad \text{using independence} \\ &= (1 - F_X(z))(1 - F_Y(z)) \\ &= (1 - F_X(z))^2, \quad \text{if they are iid} \end{aligned}$$

$$\text{Hence,} \quad F_Z(z) = 1 - (1 - F_X(z))(1 - F_Y(z))$$

- We can once again find density of Z if X, Y are continuous

- min fn is also easily generalized to n random variables
- Let $Z = \min(X_1, X_2, \cdots, X_n)$

$$\begin{aligned} P[Z > z] &= P[\min(X_1, X_2, \cdots, X_n) > z] \\ &= P[X_1 > z, \cdots, X_n > z] \\ &= P[X_1 > z] \cdots P[X_n > z], \quad \text{using independence} \\ &= (1 - F_{X_1}(z)) \cdots (1 - F_{X_n}(z)) \\ &= (1 - F_X(z))^n, \quad \text{if they are iid} \end{aligned}$$

- Hence, when X_i are iid, the df of Z is

$$F_Z(z) = 1 - (1 - F_X(z))^n$$

where F_X is the common df

Sum of two discrete rv's

- Let $X, Y \in \{0, 1, \cdots\}$
- Let $Z = X + Y$. Then we have

$$f_Z(z) = P[X + Y = z] = \sum_{\substack{x,y:\\ x+y=z}} P[X = x, Y = y]$$

$$= \sum_{k=0}^{\infty} P[X = k, Y = z - k] = \sum_{k=0}^{z} P[X = k, Y = z - k]$$

$$= \sum_{k=0}^{z} f_{XY}(k, z - k)$$

- Now suppose X, Y are independent. Then

$$f_Z(z) = \sum_{k=0}^{z} f_X(k) f_Y(z - k)$$

- Now suppose X, Y are independent Poisson with parameters λ_1, λ_2. And, $Z = X + Y$.

$$
\begin{aligned}
f_Z(z) &= \sum_{k=0}^{z} f_X(k) f_Y(z-k) \\
&= \sum_{k=0}^{z} \frac{\lambda_1^k}{k!} e^{-\lambda_1} \ \frac{\lambda_2^{z-k}}{(z-k)!} e^{-\lambda_2} \\
&= e^{-(\lambda_1+\lambda_2)} \ \frac{1}{z!} \sum_{k=0}^{z} \frac{z!}{k!(z-k)!} \ \lambda_1^k \ \lambda_2^{z-k} \\
&= e^{-(\lambda_1+\lambda_2)} \ \frac{1}{z!} (\lambda_1 + \lambda_2)^z
\end{aligned}
$$

- Z is Poisson with parameter $\lambda_1 + \lambda_2$

Independence of functions of random variable

- Suppose X and Y are independent.
- Then $g(X)$ and $h(Y)$ are independent
- This is easily generalized to functions of multiple random variables.
- That is, suppose $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ are independent.
- Then, $g(X_1, \cdots, X_m)$ is independent of $h(Y_1, \cdots, Y_n)$.

- Let X_1, X_2, X_3 be independent Poisson rv (with parameters $\lambda_1, \lambda_2, \lambda_3$).
- $Z = X_1 + X_2 + X_3$.
- Can we find pmf of Z?
- Let $W = X_1 + X_2$.
  We know its pmf (Poisson with $\lambda_1 + \lambda_2$).
- Now, $Z = W + X_3$ and W and X_3 are independent.
- So, pmf of Z is Poisson with parameter $\lambda_1 + \lambda_2 + \lambda_3$.

- In a similar way we can handle other functions (of discrete rv).
- Let $Z = X - Y$. Then

$$f_Z(z) = P[X-Y = z] = \sum_y P[Y = y, X = z+y] = \sum_y f_{XY}(z+y, y)$$

- Let $Z = XY$. Then $P[Z = 0] = P[X = 0 \text{ or } Y = 0] = \sum_{x \neq 0} f_{XY}(x, 0) + \sum_{y \neq 0} f_{XY}(0, y) + f_{XY}(0, 0)$
- For $z \neq 0$ we have

$$f_Z(z) = P[XY = z] = \sum_{x \neq 0} P[X = x, Y = z/x] = \sum_{x \neq 0} f_{XY}(x, z/x)$$

- We next look at finding density for functions of continuous rv.
- We state a general theorem that is quite useful in dealing with functions of multiple random variables.
- This result is only for continuous random variables.

- Let $X_1, \cdots, X_n$ be continuous random variables with joint density $f_{X_1 \cdots X_n}$. We define $Y_1, \cdots Y_n$ by

$$Y_1 = g_1(X_1, \cdots, X_n) \quad \cdots \quad Y_n = g_n(X_1, \cdots, X_n)$$

  We think of g_i as components of $g : \Re^n \to \Re^n$.
- We assume g is continuous with continuous first partials and is invertible.
- Let h be the inverse of g. That is

$$X_1 = h_1(Y_1, \cdots, Y_n) \quad \cdots \quad X_n = h_n(Y_1, \cdots, Y_n)$$

- Each of g_i, h_i are $\Re^n \to \Re$ functions and we can write them as

$$y_i = g_i(x_1, \cdots, x_n); \quad \cdots \quad x_i = h_i(y_1, \cdots, y_n)$$

  We denote the partial derivatives of these functions by $\frac{\partial x_i}{\partial y_j}$ etc.

- The jacobian of the inverse transformation is

$$J = \frac{\partial(x_1, \cdots, x_n)}{\partial(y_1, \cdots, y_n)} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

- We assume that J is non-zero in the range of the transformation
- **Theorem**: Under the above conditions, we have

$$f_{Y_1 \cdots Y_n}(y_1, \cdots, y_n) = |J| f_{X_1 \cdots X_n}(h_1(y_1, \cdots, y_n), \cdots, h_n(y_1, \cdots, y_n))$$

Or, more compactly, $\quad f_{\mathbf{Y}}(\mathbf{y}) = |J| f_{\mathbf{X}}(h(\mathbf{y}))$

Illustration of the theorem

- Let X_1, X_2 have a joint density, $f_{X_1 X_2}$. Consider

$$\begin{aligned} Y_1 &= g_1(X_1, X_2) = X_1 + X_2 \quad (g_1(a,b) = a + b) \\ Y_2 &= g_2(X_1, X_2) = X_1 - X_2 \quad (g_2(a,b) = a - b) \end{aligned}$$

  This transformation is invertible

$$\begin{aligned} X_1 &= h_1(Y_1, Y_2) = \frac{Y_1 + Y_2}{2} \quad (h_1(a,b) = (a + b)/2) \\ X_2 &= h_2(Y_1, Y_2) = \frac{Y_1 - Y_2}{2} \quad (h_2(a,b) = (a - b)/2) \end{aligned}$$

  The jacobian is: $\begin{vmatrix} 0.5 & 0.5 \\ 0.5 & -0.5 \end{vmatrix} = -0.5.$

- This gives: $f_{Y_1 Y_2}(y_1, y_2) = 0.5\ f_{X_1 X_2}\left(\frac{y_1+y_2}{2},\ \frac{y_1-y_2}{2}\right)$

Density of $X_1 + X_2$

- Let X, Y have joint density f_{XY}. Let $Z = X + Y$.
- We want to find f_Z using the theorem.
- To use the theorem, we need an invertible transformation of $\Re^2$ onto $\Re^2$ of which one component is $x + y$.
  We are free to choose the other function.
- We can take $Z = X + Y$ and $W = X - Y$.
- Hence we get

$$f_{ZW}(z, w) = \frac{1}{2} f_{XY}\left(\frac{z + w}{2}, \frac{z - w}{2}\right)$$

- Now we get density of Z as

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{2} f_{XY}\left(\frac{z + w}{2}, \frac{z - w}{2}\right) \, dw$$

- $Z = X + Y$ and $W = X - Y$. Then

$$
\begin{aligned}
f_Z(z) &= \int_{-\infty}^{\infty} \frac{1}{2} f_{XY}\left(\frac{z+w}{2}, \frac{z-w}{2}\right)\, dw \\
& \qquad \text{change the variable: } \; t = \frac{z+w}{2} \quad \Rightarrow dt = \frac{1}{2}\, dw \\
& \qquad\qquad\qquad \Rightarrow \; w = 2t - z \; \Rightarrow z - w = 2z - 2t \\
f_Z(z) &= \int_{-\infty}^{\infty} f_{XY}(t, z-t)\, dt \\
&= \int_{-\infty}^{\infty} f_{XY}(z-s, s)\, ds,
\end{aligned}
$$

- If, X, Y are independent

$$
f_Z(z) = \int_{-\infty}^{\infty} f_X(t)\, f_Y(z-t)\, dt
$$

- let $Z = X + Y$ and $W = X - Y$. We got

$$f_{ZW}(z, w) = \frac{1}{2} f_{XY}\left(\frac{z+w}{2}, \frac{z-w}{2}\right)$$

- Now we can calculate f_W also.

$$f_W(w) = \int_{-\infty}^{\infty} \frac{1}{2} f_{XY}\left(\frac{z+w}{2}, \frac{z-w}{2}\right)\, dz$$

$$\text{change the variable: } t = \frac{z+w}{2} \quad \Rightarrow dt = \frac{1}{2}\, dz$$

$$\Rightarrow \quad z = 2t - w \quad \Rightarrow z - w = 2t - 2w$$

$$f_W(w) = \int_{-\infty}^{\infty} f_{XY}(t, t-w)\, dt$$

$$= \int_{-\infty}^{\infty} f_{XY}(s+w, s) ds,$$

Example

- Let X, Y be iid $U[0,\ 1]$. Let $Z = X - Y$.

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(t)\ f_Y(t-z)\ dt$$

- For the integrand to be non-zero
  - $0 \leq t \leq 1 \ \Rightarrow \ \ t \geq 0, \ \ t \leq 1$
  - $0 \leq t - z \leq 1 \ \Rightarrow \ \ t \geq z, \ \ t \leq 1 + z$
  - $\Rightarrow \ \ \max(0,\ z) \leq t \leq \min(1,\ 1+z)$
- Thus, we get density as (note $Z \in (-1, 1)$)

$$f_Z(z) = \begin{cases} \int_0^{1+z} 1\ dt = 1 + z, & \text{if } -1 \leq z \leq 0 \\ \int_z^1 1\ dt = 1 - z, & 0 \leq z \leq 1 \end{cases}$$

- Thus, when $X, Y \sim U(0,1)$ iid

$$f_{X-Y}(z) = 1 - |z|, \quad -1 < z < 1$$

Sums of independent continuous rv

- Recall that *Gamma*(α, λ) density is

$$f(x) = \frac{1}{\Gamma(\alpha)} \; (\lambda)^{\alpha} \, x^{\alpha-1} \, e^{-\lambda x}, \;\; x > 0$$

- If $X \sim Gamma(\alpha_1, \lambda)$ and $Y \sim Gamma(\alpha_2, \lambda)$, and X, Y independent, then
  $X + Y \sim Gamma(\alpha_1 + \alpha_2, \lambda)$
- Similarly, sum of independent Gaussians is Gaussian
- If $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and X, Y independent, then
  $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Products and quotients of random variables

- We can use similar method for products and quotients.
- Suppose we want density of XY
- We can choose: $Z = XY \quad W = Y$
  This is invertible: $X = Z/W \quad Y = W$
- Suppose we want density of X/Y.
- We can choose: $Z = X/Y \quad W = Y$
  This is invertible: $X = ZW \quad Y = W$

Densities of standard functions of rv's

- Densities of sum, difference, product and quotient of two random variables.

$$
\begin{aligned}
f_{X+Y}(z) &= \int_{-\infty}^{\infty} f_{XY}(t, z-t)\, dt = \int_{-\infty}^{\infty} f_{XY}(z-t, t)\, dt \\
f_{X-Y}(z) &= \int_{-\infty}^{\infty} f_{XY}(t, t-z)\, dt = \int_{-\infty}^{\infty} f_{XY}(t+z, t) dt
\end{aligned}
$$

$$f_{X*Y}(z) = \int_{-\infty}^{\infty} \left|\frac{1}{t}\right| \, f_{XY}\left(\frac{z}{t}, t\right) \, dt = \int_{-\infty}^{\infty} \left|\frac{1}{t}\right| \, f_{XY}\left(t, \frac{z}{t}\right) \, dt$$

$$f_{(X/Y)}(z) = \int_{-\infty}^{\infty} |t| \, f_{XY}(zt, t) \, dt = \int_{-\infty}^{\infty} \left|\frac{t}{z^2}\right| \, f_{XY}\left(t, \frac{t}{z}\right) \, dt$$

Expectation of functions of multiple rv

- **Theorem**: Let $Z = g(X_1, \cdots X_n) = g(\mathbf{X})$. Then

$$E[Z] = \int_{\Re^n} g(\mathbf{x}) \; dF_{\mathbf{X}}(\mathbf{x})$$

- That is, if they have a joint density, then

$$E[Z] = \int_{\Re^n} g(\mathbf{x}) \; f_{\mathbf{X}}(\mathbf{x}) \; d\mathbf{x}$$

- Similarly, if all X_i are discrete

$$E[Z] = \sum_{\mathbf{x}} g(\mathbf{x}) \; f_{\mathbf{X}}(\mathbf{x})$$

- Let $Z = X + Y$. Let X, Y have joint density f_{XY}

$$\begin{aligned}
E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)\, f_{XY}(x,y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{XY}(x,y)\, dy\, dx \\
&\qquad + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} x\, f_X(x)\, dx + \int_{-\infty}^{\infty} y\, f_Y(y)\, dy \\
&= E[X] + E[Y]
\end{aligned}$$

- Expectation is a linear operator.
- This is true for all random variables.

- We saw $E[X+Y] = E[X] + E[Y]$.
- Let us calculate $\text{Var}(X+Y)$.

$$\begin{aligned}
\text{Var}(X+Y) &= E\left[\left(\,(X+Y) - E[X+Y]\,\right)^2\right] \\
&= E\left[\left(\;(X-EX) + (Y-EY)\;\right)^2\right] \\
= E\,\Big[\,(X-EX)^2 \;+\; (Y-EY)^2 \;+\; & 2(X-EX)(Y-EY)\,\Big] \\
&= E\left[(X-EX)^2\right] + E\left[(Y-EY)^2\right] \\
&\quad + 2E\left[(X-EX)(Y-EY)\right] \\
&= \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X,Y)
\end{aligned}$$

where we define **covariance** between X, Y as

$$\text{Cov}(X,Y) = E\left[(X-EX)(Y-EY)\right]$$

- We define **covariance** between X and Y by

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E\left[(X - EX)(Y - EY)\right] \\ &= E\left[XY - X(EY) - Y(EX) + EX\ EY\right] \\ &= E[XY] - EX\ EY\end{aligned}$$

- Note that $\mathrm{Cov}(X, Y)$ can be positive or negative
- We have

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\ \mathrm{Cov}(X, Y)$$

- X and Y are said to be uncorrelated if $\mathrm{Cov}(X, Y) = 0$
- If X and Y are uncorrelated then

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

- Note that $E[X + Y] = E[X] + E[Y]$ for all random variables.

Example

- Consider the joint density

$$f_{XY}(x,y) = 2, \;\; 0 < x < y < 1$$

- We want to calculate $\text{Cov}(X,Y)$

$$EX = \int_0^1 \int_x^1 x\;2\;dy\;dx = 2\int_0^1 x\;(1-x)\;dx = \frac{1}{3}$$

$$EY = \int_0^1 \int_0^y y\;2\;dx\;dy = 2\int_0^1 y^2\;dy = \frac{2}{3}$$

$$E[XY] = \int_0^1 \int_0^y xy\;2\;dx\;dy = 2\int_0^1 y\;\frac{y^2}{2}\;dy = \frac{1}{4}$$

- Hence, $\text{Cov}(X,Y) = E[XY] - EX\;EY = \frac{1}{4} - \frac{2}{9} = \frac{1}{36}$

Independent random variables are uncorrelated

- Suppose X, Y are independent. Then

$$\begin{aligned} E[XY] &= \int\int x\ y\ f_{XY}(x,y)\ dx\ dy \\ &= \int\int x\ y\ f_X(x)\ f_Y(y)\ dx\ dy \\ &= \int x f_X(x)\ dx \int y f_Y(y)\ dy = EX\ EY \end{aligned}$$

- Then, $\text{Cov}(X, Y) = E[XY] - EX\ EY = 0$.
- X, Y independent $\Rightarrow$ X, Y uncorrelated

Uncorrelated random variables may not be independent

- Suppose $X \sim \mathcal{N}(0,1)$ Then, $EX = EX^3 = 0$
- Let $Y = X^2$ Then,

$$E[XY] = EX^3 = 0 = EX\ EY$$

- Thus X, Y are uncorrelated.
- Are they independent? No
  e.g.,

$$P[X > 2 \,|\, Y < 1] = 0 \neq P[X > 2]$$

- X, Y are uncorrealted does not imply they are independent.

- We define the **correlation coefficient** of X, Y by

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

- If X, Y are uncorrelated then $\rho_{XY} = 0$.
- We can show that $|\rho_{XY}| \leq 1$
- Hence $-1 \leq \rho_{XY} \leq 1, \ \forall X, Y$
- $|\rho_{XY}| = 1$ only when $Y = aX$.

- We have $E\left[(\alpha X+\beta Y)^2\right] \geq 0, \ \forall \alpha, \beta \in \Re$

$$\begin{aligned}
\alpha^2 E[X^2] + \beta^2 E[Y^2] + 2\alpha\beta E[XY] &\geq 0, \quad \forall \alpha, \beta \in \Re \\
\text{Take} \quad \alpha = -\frac{E[XY]}{E[X^2]} & \\
\frac{(E[XY])^2}{E[X^2]} + \beta^2 E[Y^2] - 2\beta \frac{(E[XY])^2}{E[X^2]} &\geq 0, \quad \forall \beta \in \Re \\
a\beta^2 + b\beta + c \geq 0, \ \forall \beta \ \Rightarrow b^2 - 4ac \leq 0 & \\
\Rightarrow \ 4\left(\frac{(E[XY])^2}{E[X^2]}\right)^2 - 4E[Y^2]\frac{(E[XY])^2}{E[X^2]} &\leq 0 \\
\Rightarrow \ \left(\frac{(E[XY])^2}{E[X^2]}\right)^2 \leq \frac{E[Y^2](E[XY])^2}{E[X^2]} & \\
\Rightarrow \ \frac{(E[XY])^4}{(E[XY])^2} \leq \frac{E[Y^2](E[X^2])^2}{E[X^2]} & \\
\Rightarrow \ (E[XY])^2 \leq E[X^2]E[Y^2] &
\end{aligned}$$

- We showed that
$$(E[XY])^2 \leq E[X^2]E[Y^2]$$
- Take $X - EX$ in place of X and $Y - EY$ in place of Y in the above algebra.
- This gives us
$$(E[(X - EX)(Y - EY)])^2 \leq E[(X - EX)^2]E[(Y - EY)^2]$$
$$\Rightarrow \quad (\text{Cov}(X, Y))^2 \leq \text{Var}(X)\text{Var}(Y)$$
- Hence we get
$$\rho_{XY}^2 = \left(\frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \right)^2 \leq 1$$
- The equality holds here only if $E\left[(\alpha X + \beta Y)^2\right] = 0$
$$\text{Thus,} \quad |\rho_{XY}| = 1 \;\text{ only if }\; \alpha X + \beta Y = 0$$
- Correlation coefficient of X, Y is ± 1 only when Y is a linear function of X

- The covariance of X, Y is

  $$\text{Cov}(X, Y) = E[(X - EX)\,(Y - EY)] = E[XY] - EX\;EY$$

  Note that $\text{Cov}(X, X) = \text{Var}(X)$
- X, Y are called uncorrelated if $\text{Cov}(X, Y) = 0$.
- X, Y independent $\Rightarrow$ X, Y uncorrelated.
- Uncorrelated random variables need not necessarily be independent
- Covariance plays an important role in linear least squares estimation.
- Informally, covariance captures the 'linear dependence' between the two random variables.

Covariance Matrix

- Let $X_1, \cdots, X_n$ be random variables (on the same probability space)
- We represent them as a vector $\mathbf{X}$.
- As a notation, all vectors are column vectors: $\mathbf{X} = (X_1, \cdots, X_n)^T$
- We denote $E[\mathbf{X}] = (EX_1, \cdots, EX_n)^T$
- The $n \times n$ matrix whose $(i, j)^{th}$ element is $\text{Cov}(X_i, X_j)$ is called the covariance matrix (or variance-covariance matrix) of $\mathbf{X}$. Denoted as $\Sigma_{\mathbf{X}}$ or Σ_X

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \vdots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \text{Cov}(X_n, X_n) \end{bmatrix}$$

Covariance matrix

- If $\mathbf{a} = (a_1, \cdots, a_n)^T$ then
  $\mathbf{a}\,\mathbf{a}^T$ is a $n \times n$ matrix whose $(i,j)^{th}$ element is $a_i a_j$.
- Hence we get

$$\Sigma_{\mathbf{X}} = E\left[(\mathbf{X} - E\mathbf{X})\,(\mathbf{X} - E\mathbf{X})^T\right]$$

- This is because
  $\left((\mathbf{X} - E\mathbf{X})\,(\mathbf{X} - E\mathbf{X})^T\right)_{ij} = (X_i - EX_i)(X_j - EX_j)$
  and $(\Sigma_{\mathbf{X}})_{ij} = E[(X_i - EX_i)(X_j - EX_j)]$

- Recall the following about vectors and matrices
- let $\mathbf{a}, \mathbf{b} \in \Re^n$ be column vectors. Then

$$\left(\mathbf{a}^T\mathbf{b}\right)^2 = \left(\mathbf{a}^T\mathbf{b}\right)^T \left(\mathbf{a}^T\mathbf{b}\right) = \mathbf{b}^T\mathbf{a}\,\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\left(\mathbf{a}\,\mathbf{a}^T\right)\mathbf{b}$$

- Let A be an $n \times n$ matrix with elements a_{ij}. Then

$$\mathbf{b}^T A\mathbf{b} = \sum_{i,j=1}^{n} b_i b_j a_{ij}$$

  where $\mathbf{b} = (b_1, \cdots, b_n)^T$
- A is said to be positive semidefinite if $\mathbf{b}^T A\mathbf{b} \geq 0, \ \forall \mathbf{b}$

- Σ_X is a real symmetric matrix
- Let $\mathbf{a} \in \Re^n$ and let $Y = \mathbf{a}^T\mathbf{X} = \sum_i a_i X_i$.
- Then, $EY = \sum_i a_i EX_i = \mathbf{a}^T E\mathbf{X}$.
  We get variance of Y as

$$
\begin{aligned}
\text{Var}(Y) &= E[(Y - EY)^2] = E\left[\left(\mathbf{a}^T\mathbf{X} - \mathbf{a}^T E\mathbf{X}\right)^2\right] \\
&= E\left[\left(\mathbf{a}^T(\mathbf{X} - E\mathbf{X})\right)^2\right] \\
&= E\left[\mathbf{a}^T(\mathbf{X} - E\mathbf{X})\,(\mathbf{X} - E\mathbf{X})^T\mathbf{a}\right] \\
&= \mathbf{a}^T E\left[(\mathbf{X} - E\mathbf{X})\,(\mathbf{X} - E\mathbf{X})^T\right]\mathbf{a} \\
&= \mathbf{a}^T\Sigma_X\mathbf{a}
\end{aligned}
$$

- This gives $\mathbf{a}^T\Sigma_X\mathbf{a} \geq 0, \ \forall\mathbf{a}$
- This shows Σ_X is positive semidefinite

- $Y = \mathbf{a}^T \mathbf{X} = \sum_i a_i X_i$ – linear combination of X_i's.
- We know how to find its mean and variance

$$\begin{aligned} EY &= \mathbf{a}^T E\mathbf{X} = \sum_i a_i EX_i; \\ \text{Var}(Y) &= \mathbf{a}^T \Sigma_X \mathbf{a} = \sum_{i,j} a_i a_j \text{Cov}(X_i, X_j) \end{aligned}$$

- Specifically, by taking all components of $\mathbf{a}$ to be 1, we get

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i,j=1}^n \text{Cov}(X_i, X_j) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i=1}^n \sum_{j \neq i} \text{Cov}(X_i, X_j)$$

- If X_i are uncorrelated, variance of sum is sum of variances.

- Covariance matrix Σ_X positive semidefinite because

$$\mathbf{a}^T \Sigma_X \, \mathbf{a} = \text{Var}(\mathbf{a}^T \mathbf{X}) \geq 0$$

- Σ_X would be positive definite if $\mathbf{a}^T \Sigma_X \, \mathbf{a} > 0, \; \forall \mathbf{a} \neq 0$
- It would fail to be positive definite if $\text{Var}(\mathbf{a}^T \mathbf{X}) = 0$ for some nonzero $\mathbf{a}$.
- $\text{Var}(Z) = E[(Z - EZ)^2] = 0$ implies $Z = EZ$, a constant.
- Hence, Σ_X fails to be positive definite only if there is a non-zero linear combination of X_i's that is a constant.

Joint moments

- Given two random variables, X, Y
- The joint moment of order (i, j) is defined by

$$m_{ij} = E[X^i Y^j]$$

$m_{10} = EX$, $m_{01} = EY$, $m_{11} = E[XY]$ and so on

- Similarly joint central moments of order (i, j) are defined by

$$s_{ij} = E\left[(X - EX)^i (Y - EY)^j\right]$$

$s_{10} = s_{01} = 0$, $s_{11} = \text{Cov}(X, Y)$, $s_{20} = \text{Var}(X)$ and so on

- We can similarly define joint moments of multiple random variables

- We can define moment generating function of X, Y by

$$M_{XY}(s,t) = E\left[e^{sX+tY}\right], \quad s,t \in \Re$$

- This is easily generalized to n random variables

$$M_{\mathbf{X}}(\mathbf{s}) = E\left[e^{\mathbf{s}^T\mathbf{X}}\right], \quad \mathbf{s} \in \Re^n$$

- Once again, we can get all the moments by differentiating the moment generating function

$$\left.\frac{\partial}{\partial s_i} M_{\mathbf{X}}(\mathbf{s})\right|_{\mathbf{s}=0} = EX_i$$

- More generally

$$\left.\frac{\partial^{m+n}}{\partial s_i^n \, \partial s_j^m} M_{\mathbf{X}}(\mathbf{s})\right|_{\mathbf{s}=0} = EX_i^n X_j^m$$

Conditional Expectation

- Suppose X, Y have a joint density f_{XY}
- Consider the conditional density $f_{X|Y}(x|y)$. This is a density in x for every value of y.
- Since it is a density, we can use it in an expectation integral: $\int g(x)\, f_{X|Y}(x|y)\, dx$
- This is like expectation of $g(X)$ since $f_{X|Y}(x|y)$ is a density in x.
- However, its value would be a function of y.
- That is, this is a kind of expectation that is a function of Y (and hence is a random variable)
- It is called conditional expectation.

- Let X, Y be discrete random variables (on the same probability space).
- The conditional expectation of $h(X)$ conditioned on Y is a function of Y, and is defined by $E[h(X)|Y] = g(Y)$ where

$$E[h(X)|Y = y] = g(y) = \sum_x h(x)\, f_{X|Y}(x|y)$$

- Thus

$$\begin{aligned} E[h(X)|Y = y] &= \sum_x h(x)\, f_{X|Y}(x|y) \\ &= \sum_x h(x)\, P[X = x|Y = y] \end{aligned}$$

- Note that, $E[h(X)|Y]$ is a random variable

- Let X, Y have joint density f_{XY}.
- The conditinal expectation of $h(X)$ conditioned on Y is a function of Y, and its value for any y is defined by

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

- Once again, what this means is that $E[h(X)|Y] = g(Y)$ where

$$g(y) = \int_{-\infty}^{\infty} h(x)\ f_{X|Y}(x|y)\ dx$$

$E[h(X)|Y=y] = \sum_x h(x)\ f_{X|Y}(x|y)$

- Let $X = a_1 I_A + a_2 I_{A^c}$. Then $EX = a_1 P(A) + a_2 P(A^c)$
- Let $Y = b_1 I_B + b_2 I_{B^c}$. Then

$$\begin{aligned} E[X \mid Y = b_1] &= a_1 f_{X|Y}(a_1|b_1) + a_2 f_{X|Y}(a_2|b_1) \\ &= a_1 P[X = a_1|Y = b_1] + a_2 P[X = a_2|Y = b_1] \\ &= a_1 P(A|B) + a_2 P(A^c|B) \end{aligned}$$

similarly

$$E[X \mid Y = b_2] = a_1 P(A|B^c) + a_2 P(A^c|B^c)$$

- $E[X|Y]$ can be thought of as 'partial averaging' (based on Y)

A simple example

- Consider the joint density

$$f_{XY}(x, y) = 2, \;\; 0 < x < y < 1$$

- We calculated the conditional densities earlier

$$f_{X|Y}(x|y) = \frac{1}{y}, \; 0 < x < y < 1$$

- Now we can calculate the conditional expectation

$$\begin{aligned} E[X|Y = y] &= \int_{-\infty}^{\infty} x \; f_{X|Y}(x|y) \; dx \\ &= \int_0^y x \; \frac{1}{y} \; dx = \frac{1}{y} \;\; \frac{x^2}{2}\bigg|_0^y = \frac{y}{2} \end{aligned}$$

- This gives: $E[X|Y] = \frac{Y}{2}$

A simple example

- For the joint density

$$f_{XY}(x, y) = 2, \;\; 0 < x < y < 1$$

- the other conditional density is

$$f_{Y|X}(y|x) = \frac{1}{1-x}, \; 0 < x < y < 1$$

- Now we can once again calculate the conditional expectation

$$\begin{aligned} E[Y|X = x] &= \int_{-\infty}^{\infty} y \; f_{Y|X}(y|x) \; dy \\ &= \int_{x}^{1} y \; \frac{1}{1-x} \; dy = \frac{1+x}{2} \end{aligned}$$

- This gives $E[Y|X] = \frac{1+X}{2}$

- The conditional expectation is defined by

$$E[h(X)|Y=y] \;=\; \sum_x h(x)\, f_{X|Y}(x|y), \quad X, Y \text{ are discrete}$$

$$E[h(X)|Y=y] \;=\; \int_{-\infty}^{\infty} h(x)\, f_{X|Y}(x|y)\, dx, \quad X, Y \text{ have joint density}$$

- We can actually define $E[h(X,Y)|Y]$ also as above. That is,

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

- It has all the properties of expectation:
  1. $E[a|Y] = a$ where a is a constant
  2. $E[ah_1(X) + bh_2(X)|Y] = aE[h_1(X)|Y] + bE[h_2(X)|Y]$
  3. $h_1(X) \geq h_2(X) \;\Rightarrow\; E[h_1(X)|Y] \geq E[h_2(X)|Y]$

- Conditional expectation also has some extra properties which are very important
  - $E[\, E[h(X) \mid Y]\,] = E[h(X)]$
  - $E[h_1(X)h_2(Y) \mid Y] = h_2(Y)\, E[h_1(X)|Y]$
  - $E[h(X, Y) \mid Y = y] = E[h(X, y) \mid Y = y]$
- We will justify each of these.
- The last property above follows directly from the definition.

- Expectation of a conditional expectation is the unconditional expectation

$$E\,[\,E[h(X)\mid Y]\,] = E[h(X)]$$

  In the above, LHS is expectation of a function of Y.
- Let us denote $g(Y) = E[h(X)\mid Y]$. Then

$$\begin{aligned}
E\,[\,E[h(X)\mid Y]\,] &= E[g(Y)] \\
&= \int_{-\infty}^{\infty} g(y)\, f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} h(x)\, f_{X|Y}(x|y)\, dx \right) f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x)\, f_{XY}(x,y)\, dy\, dx \\
&= \int_{-\infty}^{\infty} h(x)\, f_X(x)\, dx \\
&= E[h(X)]
\end{aligned}$$

- Any factor that depends only on the conditioning variable behaves like a constant inside a conditional expectation

$$E[h_1(X)\ h_2(Y) \mid Y] = h_2(Y)\ E[h_1(X) \mid Y]$$

- Let us denote $g(Y) = E[h_1(X)\ h_2(Y) \mid Y]$

$$\begin{aligned}
g(y) &= E[h_1(X)\ h_2(Y) \mid Y = y] \\
&= \int_{-\infty}^{\infty} h_1(x)h_2(y)\ f_{X|Y}(x|y)\ dx \\
&= h_2(y)\ \int_{-\infty}^{\infty} h_1(x)\ f_{X|Y}(x|y)\ dx \\
&= h_2(y)\ E[h_1(X) \mid Y = y]
\end{aligned}$$

$$\Rightarrow E[h_1(X)\ h_2(Y) \mid Y] = g(Y) = h_2(Y)\ E[h_1(X)|Y]$$

- A very useful property of conditional expectation is $E[\, E[X|Y]\,] = E[X]$ (Assuming all expectations exist)
- We can see this in our earlier example.

$$f_{XY}(x, y) = 2, \;\; 0 < x < y < 1$$

- We easily get: $EX = \frac{1}{3}$ and $EY = \frac{2}{3}$
- We also showed $E[X|Y] = \frac{Y}{2}$

$$E[\, E[X|Y]\,] = E\left[\frac{Y}{2}\right] = \frac{1}{3} = E[X]$$

- Similarly

$$E[\, E[Y|X]\,] = E\left[\frac{1+X}{2}\right] = \frac{1}{2}\left(1 + \frac{1}{3}\right) = \frac{2}{3} = E[Y]$$

- A property of conditional expectation is

$$E[\ E[X|Y]\ ] = E[X]$$

- We assume that all three expectations exist.
- Very useful in calculating expectations

$$EX = E[\ E[X|Y]\ ] = \sum_y E[X|Y=y]\ f_Y(y) \quad \text{or} \quad \int E[X|Y=y]\ f_Y(y)\ dy$$

  This is like total prob rule for expectations.
- Can be used to calculate probabilities of events too

$$P(A) = E[I_A] = E\left[\ E\left[I_A|Y\right]\ \right]$$

Sum of random number of random variables

- Let $X_1, X_2, \cdots$ be iid rv on the same probability space. Suppose $EX_i = \mu < \infty, \ \forall i$.
- Let N be a positive integer valued rv that is independent of all X_i ($EN < \infty$)
- Let $S = \sum_{i=1}^{N} X_i$.
- We want to calculate ES.
- We can use

$$E[S] = E[\ E[S|N]\ ]$$

- We have

$$
\begin{aligned}
E[S|N = n] &= E\left[\sum_{i=1}^{N} X_i \mid N = n\right] \\
&= E\left[\sum_{i=1}^{n} X_i \mid N = n\right] \\
&\qquad \text{since } E[h(X, Y)|Y = y] = E[h(X, y)|Y = y] \\
&= \sum_{i=1}^{n} E[X_i \mid N = n] = \sum_{i=1}^{n} E[X_i] = n\mu
\end{aligned}
$$

- Hence we get

$$
E[S|N] = N\mu \quad \Rightarrow \quad E[S] = E[N]E[X_1]
$$

Wald's formula

- We took $S = \sum_{i=1}^{N} X_i$ with N independent of all X_i.
- With iid X_i, the formula $ES = EN\ EX_1$ is valid even under some dependence between N and X_i.
- Here are one version of assumptions needed.
  - A1 $E[|X_1|] < \infty$ and $EN < \infty$ (X_i iid).
  - A2 $E\left[X_n\ I_{[N \geq n]}\right] = E[X_n]P[N \geq n],\ \forall n$
- Let $S_N = \sum_{i=1}^{N} X_i$.
- Then, $ES_N = EX_1\ EN$
- Suppose the event $[N \leq n-1]$ depends only on $X_1, \cdots, X_{n-1}$.
- Such an N is called a stopping time.
- Then the event $[N \leq n-1]$ and hence its complement $[N \geq n]$ is independent of X_n and hence $A2$ holds.

- $X_1, \cdots$ iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$.
  $S_N = \sum_{i=1}^{N} X_i$, N ind of X_i.
- we can similarly calculate

$$E\left[S^2\right] = E\left[\; E\left[S^2 \mid N\right]\;\right]$$

- Using this we can calculate the variance and show that

$$\text{Var}(S) = EN\ \text{Var}(X_1) + \text{Var}(N)\ (EX_1)^2 = EN\ \sigma^2 + \text{Var}(N)\ \mu^2$$

Example

- X is number of tosses needed to get head. (Geometric rv)
- We know $E[X] = E[\,E[X|Y]\,]$ for any Y.
- Let $Y \in \{0,\ 1\}$ be outcome of first toss. (1 for head)

$$
\begin{aligned}
E[X] &= E[\,E[X|Y]\,] \\
&= E[X|Y=1]\,P[Y=1] + E[X|Y=0]\,P[Y=0] \\
&= E[X|Y=1]\,p + E[X|Y=0]\,(1-p) \\
&= 1\,p + (1+EX)(1-p) \\
\Rightarrow &\quad EX\,(1-(1-p)) = p + (1-p) \\
\Rightarrow &\quad EX\,p = 1 \\
\Rightarrow &\quad EX = \frac{1}{p}
\end{aligned}
$$

Least squares estimation

- We want to estimate Y as a function of X.
- We want an estimate with minimum mean square error.
- We want to solve (the min is over all functions g)

$$\min_g \ E\left(Y - g(X)\right)^2$$

- We want the 'best' function (linear or nonlinear)
- The solution turns out to be

$$g^*(X) = E[Y|X]$$

- Let us prove this.

- We want to show that for all g

$$E\left[\left(E[Y \mid X] - Y\right)^2\right] \leq E\left[(g(X) - Y)^2\right]$$

- We have

$$\begin{aligned}
(g(X) - Y)^2 &= \left[(g(X) - E[Y \mid X]) + (E[Y \mid X] - Y)\right]^2 \\
&= \left(g(X) - E[Y \mid X]\right)^2 + \left(E[Y \mid X] - Y\right)^2 \\
&\qquad + 2\left(g(X) - E[Y \mid X]\right)\left(E[Y \mid X] - Y\right)
\end{aligned}$$

- Now we can take expectation on both sides.
- We first show that expectation of last term on RHS above is zero.

First consider the last term

$$
\begin{aligned}
& E\Big[(g(X) - E[Y \mid X])(E[Y \mid X] - Y)\Big] \\
= \;& E\Big[\;\; E\Big\{(g(X) - E[Y \mid X])(E[Y \mid X] - Y) \mid X\Big\} \quad\Big] \\
& \qquad\qquad \text{because} \quad E[Z] = E[\,E[Z|X]\,] \\
= \;& E\Big[\;\; (g(X) - E[Y \mid X]) \;\; E\Big\{(E[Y \mid X] - Y) \mid X\Big\} \quad\Big] \\
& \qquad\qquad \text{because} \quad E[h_1(X)h_2(Z)|X] = h_1(X)\, E[h_2(Z)|X] \\
= \;& E\Big[\;\; (g(X) - E[Y \mid X]) \;\; \big(E\Big\{(E[Y \mid X])|X\Big\} \; - \; E\{Y \mid X\}\big) \quad\Big] \\
= \;& E\Big[\;\; (g(X) - E[Y \mid X]) \;\; (E[Y \mid X] \; - \; E[Y \mid X)) \quad\Big] \\
= \;& 0
\end{aligned}
$$

- We earlier got

$$
\begin{aligned}
(g(X) - Y)^2 \;&=\; \big(g(X) - E[Y \mid X]\big)^2 + \big(E[Y \mid X] - Y\big)^2 \\
&\qquad + 2\big(g(X) - E[Y \mid X]\big)\big(E[Y \mid X] - Y\big)
\end{aligned}
$$

- Hence we get

$$
\begin{aligned}
E\left[\,(g(X) - Y)^2\,\right] \;&=\; E\left[\,(g(X) - E[Y \mid X])^2\,\right] \\
&\qquad + E\left[\,(E[Y \mid X] - Y)^2\,\right] \\
&\geq\; E\left[\,(E[Y \mid X] - Y)^2\,\right]
\end{aligned}
$$

- Since the above is true for all functions g, we get

$$
g^*(X) \;=\; E\left[Y \mid X\right]
$$

Tower property of Conditional Expectation

- Conditional expectation satisfies

$$E[\ E[h(X)|Y,Z]\ |\ Y] = E[h(X)|Y]$$

  Note that all these can be random vectors.

- Let

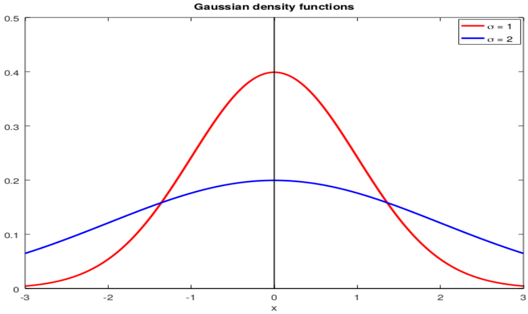
$$\begin{aligned} g_1(Y,Z) &= E[h(X)|Y,Z] \\ g_2(Y) &= E[g_1(Y,Z)|Y] \end{aligned}$$

  We can show $g_2(Y) = E[h(X)|Y]$

- Some times called the tower property of conditional expectation.

Gaussian or Normal density

- The Gaussian or normal density is given by
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$
- If X has this density, we denote it as $X \sim \mathcal{N}(\mu, \sigma^2)$.
We showed $EX = \mu$ and $\text{Var}(X) = \sigma^2$
- The density is a 'bell-shaped' curve



- Standard Normal rv — $X \sim \mathcal{N}(0,1)$
- The distribution function of standard normal is

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt$$

- Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$

$$\begin{aligned} P[a \leq X \leq b] &= P\left[\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right] \quad \text{since} \;\; \sigma > 0 \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

$$\text{because} \;\; \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1) \;\; \text{when} \;\; X \sim \mathcal{N}(\mu, \sigma^2)$$

- We can express probability of events involving all Normal rv using Φ.

- $X \sim \mathcal{N}(0,1)$. Then its mgf is

$$\begin{aligned}
M_X(t) &= E\left[e^{tX}\right] = \int_{-\infty}^{\infty} e^{tx} \, \frac{1}{\sqrt{2\pi}} \, e^{-\frac{x^2}{2}} \, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2-2tx)} \, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left((x-t)^2-t^2\right)} \, dx \\
&= e^{\frac{1}{2}t^2} \, \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2} \, dx \\
&= e^{\frac{1}{2}t^2}
\end{aligned}$$

- Now let $Y = \sigma X + \mu$. Then $Y \sim \mathcal{N}(\mu, \sigma^2)$.
  The mgf of Y is

$$\begin{aligned}
M_Y(t) &= E\left[e^{t(\sigma X+\mu)}\right] = e^{t\mu} \, E\left[e^{(t\sigma)X}\right] = e^{t\mu} \, M_X(t\sigma) \\
&= e^{\left(\mu t + \frac{1}{2}t^2\sigma^2\right)}
\end{aligned}$$

Multi-dimensional Gaussian

- The n-dimensional Gaussian density is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{|\Sigma|^{\frac{1}{2}} \, (2\pi)^{\frac{n}{2}}} \; e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}, \quad \mathbf{x} \in \Re^n$$

- $\boldsymbol{\mu} \in \Re^n$ and $\Sigma \in \Re^{n \times n}$ are parameters of the density and Σ is symmetric and positive definite.
- If $X_1, \cdots, X_n$ have the above joint density, they are said to be jointly Gaussian.
- We denote this by $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

Gaussian Vectors

- The n-dimensional Gaussian density is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{|\Sigma|^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \; e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}, \quad \mathbf{x} \in \Re^n$$

- When $\mathbf{X}$ has this joint density, we say it is a Gaussian vector.
- Same as saying $X_1, \cdots X_n$ are jointly Gaussian.

Multi-dimensional Gaussian Density

- The n-dimensional Gaussian density is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{|\Sigma|^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \; e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}, \quad \mathbf{x} \in \Re^n$$

- When $\mathbf{X}$ has this joint density, it can be shown that

$$E[\mathbf{X}] = \boldsymbol{\mu}, \quad \Sigma_X = E\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right] = \Sigma$$

$$\boldsymbol{M_X(s) = E\left[e^{s^T X}\right] = e^{s^T \mu + 0.5 s^T \Sigma s}}$$

Multi-dimensional Gaussian density

- $\mathbf{X} = (X_1, \cdots, X_n)^T$ are said to be jointly Gaussian if

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \, |\Sigma|^{\frac{1}{2}}} \; e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

- $E\mathbf{X} = \boldsymbol{\mu}$ and $\Sigma_X = \Sigma$.
- Suppose $\text{Cov}(X_i, X_j) = 0, \forall i \neq j \Rightarrow \Sigma_{ij} = 0, \forall i \neq j$.
- Then Σ is diagonal. Let $\Sigma = \text{diag}(\sigma_1^2, \cdots, \sigma_n^2)$.

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \; \sigma_1 \cdots \sigma_n} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2} = \prod_{i=1}^{n} \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2}$$

- This implies X_i are independent.
- If $X_1, \cdots, X_n$ are jointly Gaussian then uncorrelatedness implies independence.

- Let $\mathbf{X} = (X_1, \cdots, X_n)^T$ be jointly Gaussian:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \, |\Sigma|^{\frac{1}{2}}} \, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

- Let $\mathbf{Y} = \mathbf{X} - \boldsymbol{\mu}$.
- This is invertible: $\mathbf{X} = \mathbf{Y} + \boldsymbol{\mu}$. Hence

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} \, |\Sigma|^{\frac{1}{2}}} \, e^{-\frac{1}{2}\mathbf{y}^T \Sigma^{-1} \mathbf{y}}$$

- So, we can substract mean like this and work with mean zero variables.

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} \, |\Sigma|^{\frac{1}{2}}} \; e^{-\frac{1}{2}\mathbf{y}^T \Sigma^{-1} \mathbf{y}}$$

- Let L be an orthogonal matrix with $|L| = 1$, such that $L^T \Sigma^{-1} L = \text{diag}(m_1, \cdots, m_n)$. Then, $|\Sigma^{-1}| = m_1 \cdots m_n$.
- Let $\mathbf{Z} = (Z_1, \cdots, Z_n)^T = L^T \mathbf{Y}$.
- This is invertible: $\mathbf{Y} = L\mathbf{Z}$. (Because $L^{-1} = L^T$).
- Since jacobian is unity, we get

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{Y}}(L\mathbf{z}) = \frac{1}{(2\pi)^{\frac{n}{2}} \, |\Sigma|^{\frac{1}{2}}} \; e^{-\frac{1}{2}\mathbf{z}^T L^T \Sigma^{-1} L \mathbf{z}}$$

  Since the covariance matrix is now diagonal, $Z_1, \cdots, Z_n$ would be independent.
  We can see that $Z_i \sim \mathcal{N}(0, \frac{1}{m_i})$.
- If $X_1, \cdots, X_n$ are jointly Gaussian then there is a 'linear' transform that transforms them into independent random variables.

- Let X, Y be jointly Gaussian. For simplicity let $EX = EY = 0$.
- Let $\text{Var}(X) = \sigma_x^2$, $\text{Var}(Y) = \sigma_y^2$; let $\rho_{XY} = \rho \;\; \Rightarrow \;\; \text{Cov}(X, Y) = \rho\sigma_x\sigma_y$.
- Now, the covariance matrix and its inverse are given by

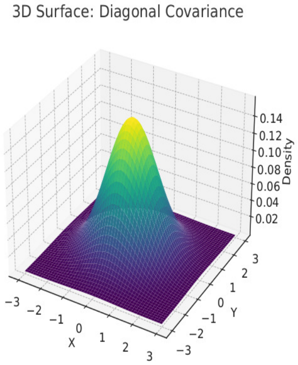
$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}; \;\; \Sigma^{-1} = \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)} \begin{bmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{bmatrix}$$

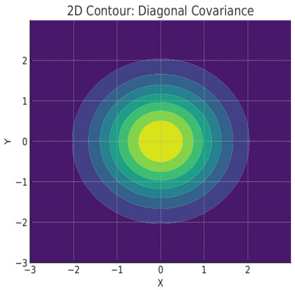
- The joint density of X, Y is given by

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \, e^{-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho xy}{\sigma_x\sigma_y}\right)}$$

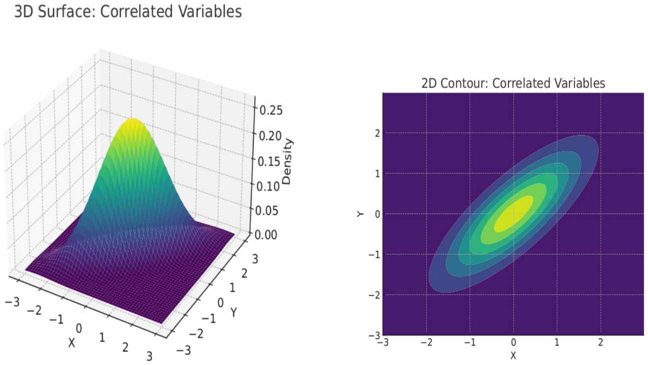
- This is the bivariate Gaussian density

- Visualization of 2D Gaussian with diagonal Σ





- 2D Gaussian where X, Y are correlated



- Suppose X, Y are jointly Gaussian (with the density above)
- Then, all the marginals and conditionals would be Gaussian.
- $X \sim \mathcal{N}(0, \sigma_x^2)$, and $Y \sim \mathcal{N}(0, \sigma_y^2)$
- $f_{X|Y}(x|y)$ would be a Gaussian density with mean $y\rho\frac{\sigma_x}{\sigma_y}$ and variance $\sigma_x^2(1-\rho^2)$.

- Let $\mathbf{X} = (X_1, \cdots, X_n)^T$ be jointly Gaussian.
- Then we call $\mathbf{X}$ as a Gaussian vector.
- It is possible that $X_i, i = 1, \cdots, n$ are individually Gaussian but $\mathbf{X}$ is not a Gaussian vector.
- Gaussian vectors have some special properties. (E.g., uncorrelated implies independence)
- Important to note that 'individually Gaussian' does not mean 'jointly Gaussian'
- Special case: If $X_1, \cdots, X_n$ are individually gaussian and independent then they are jointly Gaussian.

- The multi-dimensional Gaussian density has some important properties.
- We have seen some of them earlier.
- If $X_1, \cdots, X_n$ are jointly Gaussian then they are independent if they are uncorrelated.
- Suppose $X_1, \cdots, X_n$ be jointly Gaussian and have zero means. Then there is an orthogonal transform $\mathbf{Y} = A\mathbf{X}$ such that $Y_1, \cdots, Y_n$ are jointly Gaussian and independent.
- Another important property is the following
- $X_1, \cdots, X_n$ are jointly Gaussian if and only if $\mathbf{t}^T\mathbf{X}$ is Gaussian for for all non-zero $\mathbf{t} \in \Re^n$.
- We will prove this using moment generating functions

- Suppose $\mathbf{X} = (X_1, \cdots, X_n)^T$ be jointly Gaussian and let $W = \mathbf{t}^T\mathbf{X}$.
- Let μ_X and Σ_X denote the mean vector and covariance matrix of $\mathbf{X}$. Then

$$\mu_w \triangleq EW = \mathbf{t}^T\mu_X; \quad \sigma_w^2 \triangleq \text{Var}(W) = \mathbf{t}^T\Sigma_X\mathbf{t}$$

- The mgf of W is given by

$$\begin{aligned}
M_W(u) &= E\left[e^{uW}\right] = E\left[e^{u\,\mathbf{t}^T\mathbf{X}}\right] \\
&= M_X(u\mathbf{t}) = e^{u\mathbf{t}^T\mu_x + \frac{1}{2}u^2\mathbf{t}^T\boldsymbol{\Sigma}_x\mathbf{t}} \\
&= e^{u\mu_w + \frac{1}{2}u^2\sigma_w^2}
\end{aligned}$$

  showing that W is Gaussian
- Shows density of X_i is Gaussian for each i. For example, if we take $\mathbf{t} = (1, 0, 0, \cdots, 0)^T$ then $\mathbf{t}^T\mathbf{X}$ would be X_1.

- Now suppose $W = \mathbf{t}^T \mathbf{X}$ is Gaussian for all $\mathbf{t} \neq 0$.

$$M_W(u) = e^{u\mu_w + \frac{1}{2}u^2\sigma_w^2} = e^{u\ \mathbf{t}^T\mu_X + \frac{1}{2}u^2\ \mathbf{t}^T\boldsymbol{\Sigma}_X\mathbf{t}}$$

- This implies

$$\begin{aligned} E\left[e^{u\ \mathbf{t}^T\mathbf{X}}\right] &= e^{u\ \mathbf{t}^T\mu_X + \frac{1}{2}u^2\ \mathbf{t}^T\boldsymbol{\Sigma}_X\mathbf{t}},\ \forall u \in \Re, \forall \mathbf{t} \in \Re^n,\ \mathbf{t} \neq 0 \\ E\left[e^{\mathbf{t}^T\mathbf{X}}\right] &= e^{\mathbf{t}^T\mu_X + \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}_X\mathbf{t}},\ \forall \mathbf{t} \end{aligned}$$

  This implies $\mathbf{X}$ is jointly Gaussian.
- This is a defining property of multidimensional Gaussian density

- $\mathbf{X}$ is jointly Gaussian. Suppose A is a $k \times n$ matrix with rank k.
- Let $\mathbf{Y} = A\mathbf{X}$.
- Then Y is a Gaussian vector. (Can be shown by the same method)
- This shows all marginals of $\mathbf{X}$ are gaussian
- For example, if you take A to be

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \end{bmatrix}$$

  then $\mathbf{Y} = (X_1, X_2)^T$
- Thus marginal of any subset of the X_i would be Gaussian.

- We next consider some limits of random quantities.
- Informally, our intuition of probability of an event A is the following:
  If we independently repeat the random experiments many times, then the limit of the fraction of times A occurs would be the probability of A.
- We can rigorously formalize this notion.
- This is essentially same as the informal notion that sample mean tends to population mean in the limit.

- Let $X_1, X_2, \cdots$ be iid random variables
- Let $EX_i = \mu$ and let $\text{Var}(X_i) = \sigma^2$
- Define $S_n = \sum_{i=1}^n X_i$. Then

$$ES_n = \sum_{i=1}^n EX_i = n\mu; \quad \text{and} \quad \text{Var}(S_n) = \sum_{i=1}^n \text{Var}(X_i) = n\sigma^2$$

- Consider $\frac{S_n}{n}$, average of $X_1, \cdots, X_n$.

$$\begin{aligned} E\left[\frac{S_n}{n}\right] &= \frac{1}{n} ES_n = \mu, \ \ \forall n \\ \text{Var}\left(\frac{S_n}{n}\right) &= \left(\frac{1}{n}\right)^2 \text{Var}(S_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}, \ \ \forall n \end{aligned}$$

- X_i are iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$

$$E\left[\frac{S_n}{n}\right] = \mu; \quad \text{and} \quad \text{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}$$

- As n becomes large, variance of $\frac{S_n}{n}$ becomes close to zero
  Note that it is enough if the rv are uncorrelated for this.
- By Chebyshev Inequality

$$P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] \leq \frac{\text{Var}(\frac{S_n}{n})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}, \quad \forall \epsilon > 0$$

- Thus, we get

$$\lim_{n\to\infty} P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] = 0, \quad \forall \epsilon > 0$$

- Known as **weak law of large numbers**

- Suppose we are tossing a (biased) coin repeatedly
- Let $X_i = 1$ if i^{th} toss came up head and is zero otherwise, $i = 1, 2, \cdots$.
- $EX_i = p$ where p is the probability of heads.
- $S_n = \sum_{i=1}^{n} X_i$ is the number of heads in n tosses
- $\frac{S_n}{n}$ is the fraction of heads in n tosses.
- We are saying $\frac{S_n}{n}$ 'converges' to p
- The probability of head is the limiting fraction of heads when you toss the coin infinite times

$$\lim_{n\to\infty} P\left[\left|\frac{S_n}{n} - p\right| \geq \epsilon\right] = 0, \quad \forall \epsilon > 0$$

- This is true of any event.
- Consider repeatedly performing a random experiment
- Let X_i be the indicator of event A on i^{th} repetition
- Then $EX_i = P(A), \forall i$
- $\frac{S_n}{n}$ is the fraction of times the event A occurred.
- The fraction of times an event occurs 'converges' to its probability as you repeat the experiment infinite times

- X is a random variable and we want to find EX.
- Make multiple independent observations of X. Call them $X_1, \cdots, X_n$.
- These are called samples of X.
  Let $S_n = \sum_{i=1}^{n} X_i$
- $\frac{S_n}{n}$ is the sample mean – average of all samples.
- $\frac{S_n}{n}$ has the same expectation as X but has much smaller variance.
- Sample mean 'converges' to expectation ('population mean')
- This is the principle of sample surveys
- In general one can get an approximate value of expectation of X through simulations/experiments
- Known as Monte Carlo simulations

- X_i are iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$

$$E\left[\frac{S_n}{n}\right] = \mu; \quad \text{and} \quad \text{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}$$

- As n becomes large, variance of $\frac{S_n}{n}$ becomes close to zero
- We would like to say $\frac{S_n}{n} \to \mu$.
- We need to properly define convergence of a sequence of random variables
- One way of looking at this convergence is

$$\lim_{n\to\infty} P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] = 0, \ \ \forall \epsilon > 0$$

- There are other ways of defining convergence of random variables

Convergence in Probability

- A sequence of random variables, X_n, is said to **converge in probability** to a random variable X_0 if

$$\lim_{n \to \infty} P\left[|X_n - X_0| > \epsilon\right] = 0, \ \forall \epsilon > 0$$

  This is denoted as $X_n \stackrel{P}{\to} X_0$
- The sequence converges to a constant, c, in probability if

$$\lim_{n \to \infty} P\left[|X_n - c| > \epsilon\right] = 0, \ \forall \epsilon > 0$$

- By the definition of limit, the above means

$$\forall \delta > 0, \ \exists N < \infty, \ s.t. \ P\left[|X_n - c| > \epsilon\right] < \delta, \ \forall n > N$$

- We only need marginal distributions of individual X_n to decide whether a sequence converges to a constant in probability

Example: Partial sums of iid random variables

- X_i are iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$
- Then we saw

$$P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] \leq \frac{\sigma^2}{n\epsilon^2}, \quad \forall \epsilon > 0$$

- Hence we have $\frac{S_n}{n} \xrightarrow{P} \mu$
- Weak law of large numbers says that sample mean converges in probability to the expectation

Example

- Let $X_1, X_2, \cdots$ be a sequence of iid random variable which are uniform over $(0,\ 1)$.
- Let $M_n = \max(X_1, X_2, \cdots, X_n)$
- Does M_n converge in probability?
- A reasonable guess for the limit is 1

$$P\left[|M_n - 1| \geq \epsilon\right] = P\left[M_n \leq 1 - \epsilon\right] = (1-\epsilon)^n$$

- This implies $M_n \xrightarrow{P} 1$
- Suppose $Z_n = \min(X_1, X_2, \cdots, X_n)$.
  Then $Z_n \xrightarrow{P} 0$

Some properties of convergence in probability

- $X_n \stackrel{P}{\rightarrow} X$ and $X_n \stackrel{P}{\rightarrow} Y \;\Rightarrow\; P[X = Y] = 1$
- $X_n \stackrel{P}{\rightarrow} X \;\Rightarrow\; P[|X_n - X_m| > \epsilon] \rightarrow 0$ as $n, m \rightarrow \infty$
- Suppose $X_n \stackrel{P}{\rightarrow} X$ and $Y_n \stackrel{P}{\rightarrow} Y$ Then the following hold
    1. $aX_n \stackrel{P}{\rightarrow} aX$
    2. $X_n + Y_n \stackrel{P}{\rightarrow} X + Y$
    3. $X_n Y_n \stackrel{P}{\rightarrow} XY$
    4. $g(X_n) \stackrel{P}{\rightarrow} g(X)$ where g is a continuous function from $\Re$ to $\Re$.

Convergence in distribution

- Let F_n be the df of X_n, $n = 1, 2, \cdots$. Let X be a rv with df F.
- Sequence X_n is said to converge to X **in distribution** if

$$F_n(x) \to F(x), \ \ \forall x \ \text{ where } F \text{ is continuous}$$

- We denote this as

$$X_n \xrightarrow{d} X, \quad \text{or} \quad X_n \xrightarrow{L} X, \quad \text{or } F_n \xrightarrow{w} F$$

- This is also known as **convergence in law** or weak convergence
- Note that here we are essentially talking about convergence of distribution functions.
- Convergence in probability implies convergence in distribution
- The converse is not true. (e.g., sequence of iid random variables)

Examples

- $X_1, X_2, \cdots$ be iid; uniform over $(0,\ 1)$
- $N_n = \min(X_1, \cdots, X_n),\ Y_n = nN_n$.
  Does Y_n converge in distribution?

$$P[N_n > a] = (P[X_i > a])^n = (1-a)^n, \ \ 0 < a < 1$$

$$P[Y_n > y] = P[N_n > y/n] = \left(1 - \frac{y}{n}\right)^n, \quad \text{if} \ \ n > y$$

- Hence for any y

$$\lim_{n \to \infty} P[Y_n > y] = \lim_{n \to \infty} \left(1 - \frac{y}{n}\right)^n = e^{-y}$$

- The sequence converges in distribution to an exponential rv

- $X_n \xrightarrow{d} X$
  $\Leftrightarrow \quad F_n(x) \rightarrow F(x), \ \ \forall x$ where F is continuous
- This means that the sequence of functions F_n converge point-wise and the limit function is a distribution function.
- In general, $X_n \xrightarrow{d} X$ does not imply that the pdf's or pmf's converge point-wise to the limit pdf or pmf.
- However if the sequence of pmf's (or pdf's) converge point-wise and the limit is a pmf (or pdf) then we have $X_n \xrightarrow{d} X$.
- The following are true about convergence in distribution.
  - $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$
  - $X_n \xrightarrow{d} k \Rightarrow X_n \xrightarrow{P} k$, where k is a constant

- We have seen two different modes of convergence
- $X_n \xrightarrow{d} X$ iff

$$F_n(x) \rightarrow F(x), \ \ \forall x \ \ \text{where } F \text{ is continuous}$$

- $X_n \xrightarrow{P} X$ iff

$$\lim_{n \rightarrow \infty} P\left[|X_n - X| > \epsilon\right] = 0, \ \forall \epsilon > 0$$

- Convergence in probability implies convergence in distribution.
- There are other modes of convergence for random variables.

- An important results about sequence of independent random variables is weak law of large numbers.
- Given X_i are iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^{n} X_i$
  - Weak law of large numbers: $\frac{S_n}{n} \xrightarrow{P} \mu$
- Another useful result is the Central Limit Theorem (CLT)
- CLT is about (normalized) sums of of independent random variables converging to the Gaussian distribution

Central Limit Theorem

- Given X_i are iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $n = 1, 2, \cdots$

$$S_n = \sum_{i=1}^{n} X_i \quad E\left[S_n\right] = n\mu, \quad \text{Var}(S_n) = n\sigma^2$$

  Let

$$\tilde{S}_n = \frac{S_n - E[S_n]}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

- Central Limit Theorem states: $\tilde{S}_n \stackrel{d}{\to} \mathcal{N}(0, 1)$

Central Limit Theorem

- Given X_i iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$
- Let $\tilde{S}_n = \frac{S_n - ES_n}{\sqrt{\text{var}(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$
- **(Lindberg-Levy) Central Limit Theorem**

$$\lim_{n\to\infty} P\left[\tilde{S}_n \leq x\right] = \lim_{n\to\infty} P\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt, \quad \forall x$$

- What CLT says is that sums of iid random variables, when appropriately normalized, would always approach the Gaussian distribution.
- It allows one to approximate distribution of sums of independent rv's
- Let X_i be iid and $S_n = \sum_{i=1}^{n} X_i$

$$P[S_n \leq x] = P\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{x - n\mu}{\sigma\sqrt{n}}\right] \approx \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right)$$

- Thus, S_n is well approximated by a normal rv with mean $n\mu$ and variance $n\sigma^2$, if n is large

Example

- Twenty numbers are rounded off to the nearest integer and added. What is the probability that the sum obtained differs from true sum by more than 3.
- A reasonable assumption is round-off errors are independent and uniform over $[-0.5,\ 0.5]$
- Take $Z = \sum_{i=1}^{20} X_i$, $X_i \sim U[-0.5,\ 0.5]$, X_i iid.
- Then Z represents the error in the sum.

- $Z = \sum_{i=1}^{20} X_i$, $X_i \sim U[-0.5,\ 0.5]$, X_i iid
- $EX_i = 0$ and $\text{Var}(X_i) = \frac{1}{12}$.
- Hence, $EZ = 0$ and $\text{Var}(Z) = \frac{20}{12} = \frac{5}{3}$

$$
\begin{aligned}
P[|Z| \leq 3] &= P[-3 \leq Z \leq 3] \\
&= P\left[\frac{-3}{\sqrt{\frac{5}{3}}} \leq \frac{Z - EZ}{\sqrt{\text{Var}(Z)}} \leq \frac{3}{\sqrt{\frac{5}{3}}}\right] \\
&\approx \Phi\left(\frac{3}{\sqrt{\frac{5}{3}}}\right) - \Phi\left(\frac{-3}{\sqrt{\frac{5}{3}}}\right) \\
&\approx \Phi(2.3) - \Phi(-2.3) \\
&= 0.9893 - 0.0107 \approx 0.98
\end{aligned}
$$

- Hence probability that the sum differs from true sum by more than 3 is 0.02

- We can approximate binomial rv with Gaussian for large n
- Binomial random variable with parameters n, p is a sum of n independent Bernoulli variables:
  $S_n = \sum_{i=1}^{n} X_i; \;\; X_i \in \{0, \ 1\}, \, P[X_i = 1] = p, \, X_i$ ind
- Hence we can approximate distribution of S_n by

$$
\begin{aligned}
P[S_n \leq x] &= P\left[\frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{x - np}{\sqrt{np(1-p)}}\right] \\
&\approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right)
\end{aligned}
$$

- For large n, binomial rv is like a Gaussian rv with mean np and variance $np(1-p)$
- The approximation is quite good in practice

- S_n be binomial with parameters n, p

$$P[S_n \leq x] \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right)$$

- For example, with $p = 0.95$

$$P[S_{110} \leq 100] \approx \Phi\left(\frac{100 - 110 * 0.95}{\sqrt{110 * 0.05 * 0.95}}\right) \approx \Phi(-1.97) = 0.025$$

- Since S_n is integer-valued, the LHS above is same for all x between two consecutive integers; but RHS changes
- To get a good approximation, to calculate $P[S_n \leq m]$ one uses $P[S_n \leq m + 0.5]$ in the above approximation formula

- CLT allows one to get rate of convergence of law of large numbers
- Let X_i iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$
- By Law of large numbers, $\frac{S_n}{n} \rightarrow \mu$.
- Now, by CLT

$$
\begin{aligned}
P\left[\left|\frac{S_n}{n} - \mu\right| > \epsilon\right] &= P\left[|S_n - n\mu| > n\epsilon\right] \\
&= P\left[\left|\frac{S_n - n\mu}{\sigma\sqrt{n}}\right| > \frac{n\epsilon}{\sigma\sqrt{n}}\right] \\
&\approx 1 - \left(\Phi\left(\frac{n\epsilon}{\sigma\sqrt{n}}\right) - \Phi\left(-\frac{n\epsilon}{\sigma\sqrt{n}}\right)\right) \\
&= 2\left(1 - \Phi\left(\frac{n\epsilon}{\sigma\sqrt{n}}\right)\right)
\end{aligned}
$$

(because $\Phi(-x) = (1 - \Phi(x))$)

Example: Opinion Polls

- let p denote the fraction of population that prefers product A to product B
- We want to estimate p
- We conduct a sample survey by asking n people
- We want to make a statement such as
  $$p = 0.34 \pm 0.07 \textit{ with a confidence of } 95\%$$
- Here, the 0.34 would be the sample mean. The other two numbers can be fixed using CLT

- $X_i \in \{0, 1\}$ iid, $EX_i = p$, $S_n = \sum_{i=1}^{n} X_i$
- Now, by CLT, we have

$$P\left[\left|\frac{S_n}{n} - p\right| > \epsilon\right] = P\left[|S_n - np| > n\epsilon\right] = 2\left(1 - \Phi\left(\frac{n\epsilon}{\sqrt{np(1-p)}}\right)\right)$$

- Suppose we want to satisfy

$$P\left[\left|\frac{S_n}{n} - p\right| > \epsilon\right] = \delta$$

- We can calculate any one of ϵ, δ or n given the other two using the earlier equation.
- But we need value of p for it!

- Fortunately, $\sqrt{p(1-p)}$ does not change too much with p
- It attains its maximum value of 0.5 at $p = 0.5$
- It is 0.458 at $p = 0.3$ and is 0.4 at $p = 0.2$
- One normally fixes this variance as 0.5 or 0.45 to calculate the sample size, n.
- There are other ways of handling it

- We have

$$P\left[\left|\frac{S_n}{n} - p\right| > \epsilon\right] = 2\left(1 - \Phi\left(\frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}}\right)\right)$$

- Suppose $n = 900$ and $\epsilon = 0.025$.
Let us approximate $\sqrt{p(1-p)} = 0.45$. Then

$$2\left(1 - \Phi\left(\frac{0.025 * 30}{0.45}\right)\right) = 2(1 - \Phi(1.66)) \approx 0.1$$

- If we took $\sqrt{p(1-p)} = 0.5$ we get the value as 0.14
- If we use Chebyshev inequality with variance as 0.5 we get the bound as 0.4

A Digression – Hoeffding Bound

- In this example we saw that we need to assume something about the variance of X_i. There are other ways to handle it.
- X_i are iid rv taking values in $[a, b]$ and $S_n = \sum_{i=1}^{n} X_i$.
- Then the (two-sided) Hoeffding bound is

$$P\left[\left|\frac{S_n}{n} - p\right| > \epsilon\right] \leq 2e^{-2n\epsilon^2/(b-a)}$$

- This bound does not need any moments of X_i (but assumes they are bounded).
- When X_i are Bernoulli, $b - a = 1$.

Confidence intervals

- Let X_i iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$.
- Using CLT, we get

$$P\left[\left|\frac{S_n}{n} - \mu\right| > c\right] = 2\left(1 - \Phi\left(\frac{c\sqrt{n}}{\sigma}\right)\right)$$

- If the RHS above is δ, then we can say that $\frac{S_n}{n} \in [\mu - c,\ \mu + c]$ with probability $(1 - \delta)$
- This interval is called the $100(1-\delta)\%$ confidence interval.

$$P\left[\left|\frac{S_n}{n} - \mu\right| > c\right] = 2\left(1 - \Phi\left(\frac{c\sqrt{n}}{\sigma}\right)\right)$$

- Suppose $c = \frac{1.96\sigma}{\sqrt{n}}$
- Then

$$P\left[\left|\frac{S_n}{n} - \mu\right| > \frac{1.96\sigma}{\sqrt{n}}\right] = 2\left(1 - \Phi\left(1.96\right)\right) = 0.05$$

- Denoting $\bar{X} = \frac{S_n}{n}$, the 95% confidence interval is $\left[\bar{X} - \frac{1.96\sigma}{\sqrt{n}},\ \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right]$
- One generally uses an estimate for σ obtained from X_i
- In analyzing any experimental data the confidence intervals or the variance term is important

- $X_i, i = 1, 2, \cdots$ iid; $EX_i = 0$, $\text{Var}(X_i) = 1$.
- Let $S_n = \sum_{i=1}^{n} X_i$.
- The weak law of large numbers gives

$$\frac{S_n}{n} \xrightarrow{P} 0$$

- Central Limit theorem gives

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

central limit theorem

- CLT essentially states that sum of many independent random variables behaves like a Gaussian random variable
- It is very useful in many statistics applications.
- We stated CLT for iid random variables.
- While independence is important, all rv need not have the same distribution.
- Essentially, the variances should not die out.