Lecture 1

(31 July 2023)

Probability & Random processes

Grading Plan (Tentative)

Assignments	– 15 %
Quiz 1	– 15 %
Quiz 2	– 15 %
Mid-sem	– 20 %
End-sem	– 35 %

Module 1 (Sets and Basics of Probability).

Sets, Probabilistic models, Conditional probability, Bayes' Rule

Module 2 (Discrete Random Variables).

Module 3 (Continuous Random variables).

Module 4 (Tail Bounds and Limit Theorems).

Module 5 (Random Processes).

Textbook : Introduction to Probability

by

Bertsekas and Tsitsiklis, 2nd Edition

P. T. O

Introduction to Probability

Why study Probability:

- Randomness and uncertainity exist in our daily lives.

- Probability theory is a mathematical framework that allows us to describe and analyze random phenomena (i.e., events or experiments whose outcomes we cannot predict with certainity.

- probability (roughly) means possibility. It helps us to predict how likely or unlikely an event will occur.

[Example: Flipping a fair coin]

Different Approaches to Probability

A. Probability as the Ratio of Favourable to Total Outcomes (classical Approach)

$$\text{probability of an event } E = \frac{\text{No. of ways } E \text{ can occur}}{\text{Total no. of possible outcomes}}$$

Example. Suppose we throw a pair of unbiased dice.

1) What is the probability of getting a sum of 7?

2) What is the probability of getting a sum of 10?

Ans. 1) $6/36 = 1/6$, 2) $3/36 = 1/12$.

Example. Suppose we throw a fair coin thrice. What is the

probability of getting at most 2 Heads?

Ans. The set of possible outcomes is

{TTT, TTH, HTT, THT, THH, HTH, HHT, HHH}.

$$P(\text{at most 2 Heads}) = \frac{7}{8}$$

$$\left(= 1 - \tfrac{1}{8}\right)$$

This approach suffers from at least two significant problems.

1) It cannot deal with outcomes that are not equally likely; and

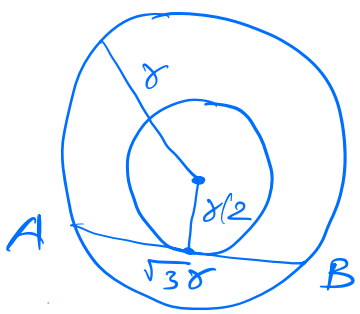
2) it cannot handle uncountably infinite outcomes without ambiguity.

(when the no. of possible outcomes is infinite, we must use length, area, or some other measure of infinity for determining the ratio)

Example (Bertrand Paradox). We are given a circle C of radius r and we wish to determine the probability p that the length l of a 'randomly selected' chord AB is greater than the length $\sqrt{3}r$ (length of the side of an equilateral triangle).

We show that this problem has at least two reasonable and different solutions.

I. If the center M of the chord AB lies inside the circle C_1 of radius $r/2$:



, then $l > \sqrt{3}r$.

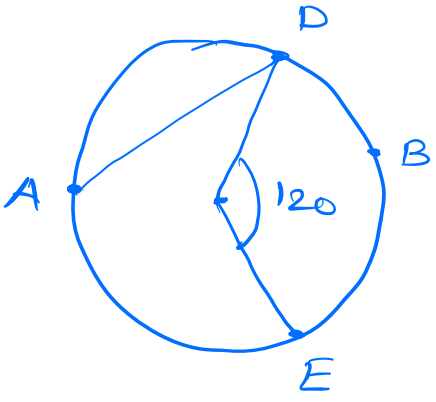
Favourable outcomes for chord center
= all points inside the circle C_1

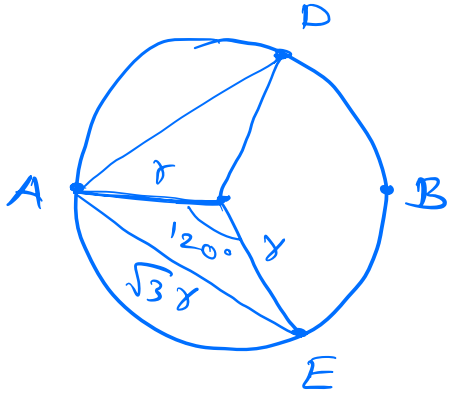
Total possible outcomes for chord center
= all points inside C

$$P = \text{area}(C_1) / \text{area}(C) = \frac{1}{4}.$$

(Using area as the measure of infinite points)

B. We assume that the end A of chord AB is fixed. This reduces the no. of possibilities but it has no effect on the value of P because the no. of favourable locations of B is reduced proportionately.





If B is on the 120° arc DBE, then $l > \sqrt{3}r$.

Favourable outcomes

= Points on this arc,

Total outcomes

= all Points on the circumference of c

$$P = \frac{2\pi r/3}{2\pi r} = \frac{1}{3}.$$

This demonstrates the ambiguities associated with the classical definition,

B. Probability as a Measure of Frequency of Occurence

- Define the probability of an event E by performing the experiment n times, No. of times E occurs is denoted by n_E.

$$P(E) = \lim_{n \to \infty} \frac{n_E}{n}.$$

clearly since $n_E \leq n$, we must have $0 \leq P(E) \leq 1$.

Disadvantages

1) We can never perform the experiment infinite number of times
2) The definition does not capture belief factor

Despite the problems with the frequency definition of probability, the relative freq. concept is essential in applying probability theory to the real world.

C. Probability Based on Axiomatic Theory

We need to introduce

- Random experiment – It is simply an experiment in which outcomes are non-deterministic that is probabilistic.
- Sample space – set of all outcomes of the experiment.
- Event – Any subset of the sample space.

Sample Space and Probability

Sets

- A set is a collection of objects, which are the elements of the set.
- A set with no elements is called the empty set, denoted by ϕ or $\{\}$.
- Finite set: A set with a finite no. of elements:

$$S = \{x_1, x_2, \ldots, x_n\}.$$

Eg. set of possible outcomes of a coin toss

$$= \{H, T\} = \{\text{heads, tails}\}.$$

set of possible outcomes of a die roll

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

- If a set contains infinitely many elements $x_1, x_2, \ldots$, which can be enumerated in a list (i.e., a bijective mapping with naturals), we write $S = \{x_1, x_2, \ldots\}$ and call S as a countably infinite set.

Eg. set of even integers $= \{0, 2, -2, 4, -4, \ldots\}$

- A set is uncountable if its elements cannot be enumerated in a list.

- Subset notation: $A \subseteq B \iff (x \in A \Rightarrow x \in B)$.
- Universal set Ω contains all objects that could be of interest in a particular context.

Set Operations.

- Complement of a set S, $S^c = \{x \in \Omega : x \notin S\}$.

$$\Omega^c = \phi, \quad \phi^c = \Omega.$$

- Union of two sets A and B

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

- Intersection of two sets A and B

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

- Infinite union.

If for every $n \in \mathbb{N}$ we are given S_n,

$$\bigcup_{n=1}^{\infty} S_n = S_1 \cup S_2 \cup \cdots = \{x \in \Omega : x \in S_n \text{ for some } n\}$$

- Infinite intersection

$$\bigcap_{n=1}^{\infty} S_n = S_1 \cap S_2 \cap \cdots = \{x \in \Omega : x \in S_n \text{ for all } n\}$$

- Two sets are said to be disjoint if their intersection is empty (i.e., there is no element in common).

  more generally given multiple no. of sets they are said to be pairwise disjoint if every pair of those sets is disjoint

- Real numbers $\mathbb{R}$

  set of pairs of reals $\mathbb{R}^2$

  set of triplets of reals $\mathbb{R}^3$.

Some Properties

- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- Demorgan's laws

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

Exercise 1.1. Prove that $\mathbb{Q} \cap [0,1]$ is a countably infinite set.

Exercise 1.2. Prove that $\{0,1\}^\infty$ is an uncountably infinite set.

[Cantor's diagonalization argument]

Exercise 1.3.

(i) For $n \in \mathbb{N}$ show that

$$\left(\bigcup_{i=1}^{n} A_i\right)^c = \bigcap_{i=1}^{n} A_i^c .$$

[Hint: Use mathematical induction]

(ii) Show that

$$\left(\bigcup_{i=1}^{\infty} S_i\right)^c = \bigcap_{i=1}^{\infty} S_i^c ,$$

$$\left(\bigcap_{i=1}^{\infty} S_i\right)^c = \bigcup_{i=1}^{\infty} S_i^c .$$

[Note that induction does not directly give such a statement for infinite number.
Reason (via a simple counter example):
Let $P(n)$ be the statement "n is finite".
$P(n)$ is true for every $n \in \mathbb{N}$,
$P(\infty)$ is false.]

Lecture 2
(3 August 2023)

Probabilistic Models

- A probabilistic model is a mathematical description of an uncertain situation or a random experiment.

Elements of a Probabilistic model:

- Sample space Ω, the set of all possible outcomes of an experiment.

  A subset of a sample space is called an event.

- Probability law which assigns a non-negative number $P(A)$ to an event A that encodes our knowledge or belief about the collective "likelihood" of the elements of A.

Sample space and Events

- Sample space is the set of all possible outcomes of a random experiment.

  [The random experiment produces exactly one out of the all the possible outcomes]

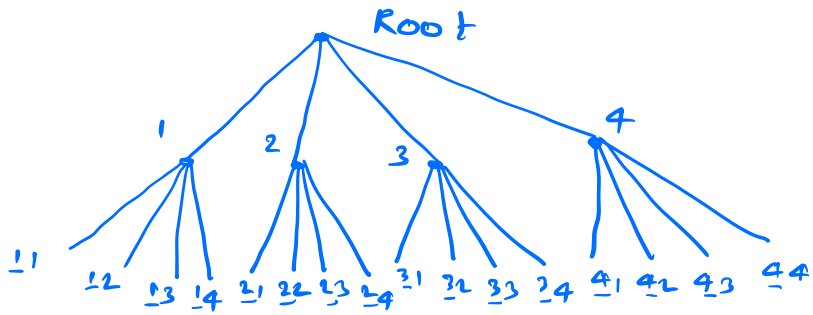
- The elements of the sample space should be
  - mutually exclusive
  - collectively exhaustive

Consider rolling a die. Is $\{1, 2, 3, 4, 5\}$ the sample space?

[No, it is not collectively exhaustive]

Example. Consider rolling a 4-sided die twice (a single random experiment).

11	12	13	14
21	22	23	24
31	32	33	34
41	42	43	44

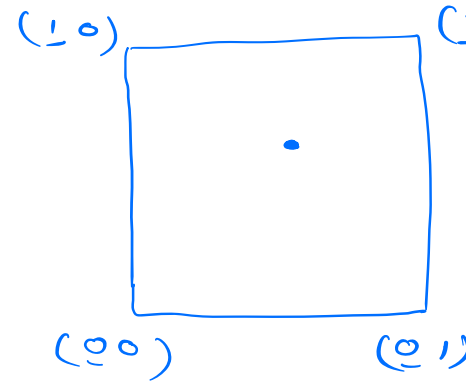


These are two equivalent descriptions of the sample space of a random experiment involving two rolls of a 4-sided die.

- Sample space of an experiment may consist of a finite or an infinite number of possible outcomes. Finite sample spaces are conceptually

and mathematically simpler, still sample spaces with an infinite number of elements are quite common,

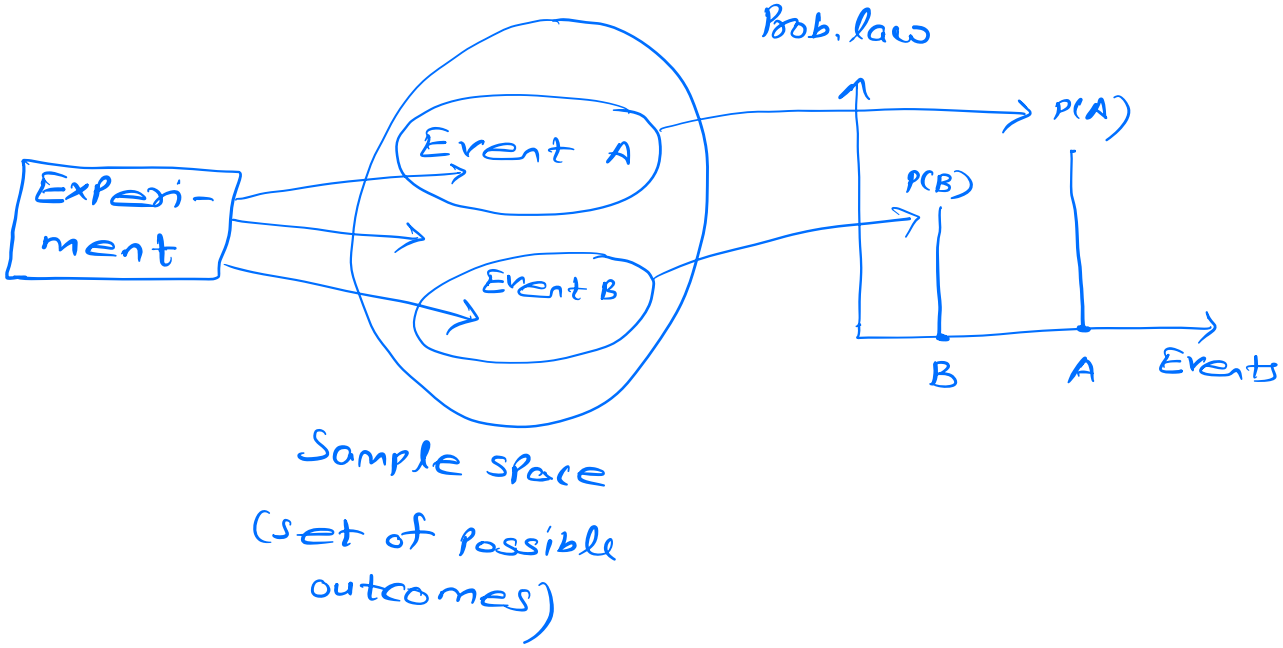
Example. Continuous sample space. Consider throwing a dart on a 1x1 square target.



Record (x,y)

$0 \le x \le 1 \quad 0 \le y \le 1$.

$\Omega = \{(x,y) : 0 \le x \le 1, \; 0 \le y \le 1\}$.



Pictorial view of a Probabilistic model

- The sample space should have enough detail to distinguish between all outcomes of interest, while avoiding the irrelevant details.

Example. Consider a single random experiment that involves 3 successive coin tosses, and there are two different scenarios of interest.

Game 1: we receive Rs. 1 each time a head comes up.

Game 2: we receive Rs. 1 for every coin toss up to and including the first time a head comes up. Then, we receive Rs. 2 for every coin toss up to the second time a head comes up, Rs. 4 up to the third time.

$\Omega_1 = \{0, 1, 2, 3\}$

$\Omega_2 = \{HHH, HHT, THH, HTH, TTH, THT, HTT, TTT\}$

1 2 4 1 2 4 1 1 2 1 2 2 1 1 1 1 2 1 2 2 1 1 1

Probability Laws

- Suppose we have settled on the sample space Ω associated with a random experiment. To complete the probabilistic model we must introduce a probability law.
- Intuitively this specifies the "likelihood" of any outcome, or any set of possible outcomes (an event). More precisely we assign a number $P(A)$ to every event A satisfying the following axioms.

Axioms of Probability:

1. (Nonnegativity) $P(A) \geq 0$ for every event A.
2. (Additivity) If A and B are two disjoint events then $P(A \cup B) = P(A) + P(B)$.
   More generally if $A_1, A_2, \ldots$ is a sequence of disjoint events, then
   $$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$
3. (Normalization) $P(\Omega) = 1$.

Properties of Probability Laws

(1) $P(A)+P(A^c)=1$.

(2) $P(A)\leq 1$.

(3) $P(\phi)=0$.

(4) For pairwise disjoint events A, B, and C,

$$P(A\cup B\cup C)=P(A)+P(B)+P(C).$$

(similarly for $n\in\mathbb{N}$ events)

(5) If $A\subseteq B$, then $P(A)\leq P(B)$.

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

(7) $P(A\cup B)\leq P(A)+P(B)$ [Union Bound].

(8) $P(A\cup B\cup C)=P(A)+P(B\cap A^c)+P(A^c\cap B^c\cap C)$.

$A\cap A^c=\phi \Rightarrow P(A\cup A^c)=P(\Omega)=1=P(A)+P(A^c)$

[This proves (1)]

$P(A)=1-P(A^c)\leq 1$ since $P(A^c)\geq 0$ by nonnegativity

[This proves (2)]

$\Omega\cap\phi=\phi \Rightarrow P(\Omega)=P(\Omega)+P(\phi)$

$\Rightarrow P(\phi)=0$. [This proves (3)]

For disjoint sets A, B, and C,

$P(A \cup B \cup C) = P(A) + P(B \cup C) = P(A) + P(B) + P(C)$

Similarly $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$ for pairwise disjoint sets $A_1, A_2, \ldots, A_n$.

[This proves (4)]

$A \subseteq B$, $B = A \cup (B \setminus A)$

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

Since A and $B \cap A^c$ are disjoint sets

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

≥ 0.

[This proves (5)]

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

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

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

$= P(A \cap B) + P(A^c \cap B)$

[This proves (6)]

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

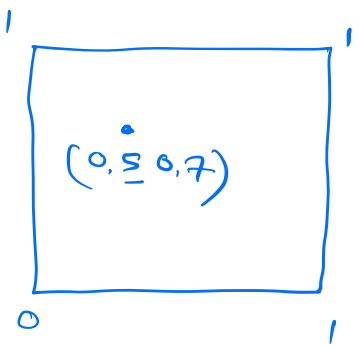
$\leq P(A) + P(B)$ ($\because$ Nonnegativity)

[This proves (7)]

$P(A \cup B) \geq \max\{P(A), P(B)\}$

Example. (Continuous model)

Probabilistic models with continuous sample spaces differ from their discrete counterparts in that the probabilities of the single-element events may not be sufficient to characterize the probability law. Consider throwing a dart on 1x1 square target.



$A \subseteq [0,1]^2$, $P(A) =$ area of A.

$$P(\{(0.5, 0.7)\}) = 0$$

$$P([0, 0.5]^2) = \frac{1}{4}$$

Lecture 3

(7 August 2023)

Conditional Probability

- Provides us with a way to reason about the outcome of an experiment based on Partial information.

- More precisely given an experiment a corresponding sample space and a probability law, suppose that we know that the outcome is within some given event B, we wish to quantify the likelihood that the outcome also belongs to some other given event A.

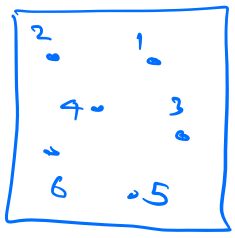
- Conditional probability of A given B denoted by $P(A|B)$.

Example.

On rolling a die what is the probability that the outcome is 2 given that the outcome is even?

Given an experiment a corresponding sample space and a probability law suppose that we know that the outcome is in B. We want to quantify the likelihood that the outcome also belongs to some other event A.

$P(\text{outcome is } 1 \mid \text{the outcome is even}) = 0$



$P(A \mid B) = 0 \quad \text{if } A \cap B = \phi$

$\propto P(A \cap B)$

$P(\text{outcome is } 2 \mid \text{the outcome is even}) = \frac{1}{3} = (1/6)/(1/2)$

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

Conditional probability of A given B s.t. $P(B) > 0$

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

- This specifies a new probability law on the same sample space Ω.

Exercise 3.1. For a fixed $B \subseteq \Omega$, prove that $P(\cdot|B)$ satisfies the probability axioms.

- If the possible outcomes are finitely many and equally likely, then

$$P(A|B) = \frac{\text{no. of elements in } A \cap B}{\text{no. of elements in } B}$$

Exercise 3.3 A fair 4-sided die is rolled twice and we assume that all sixteen possible outcomes are equally likely. Let x and y be the results of the 1st and the 2nd roll respectively. Find $P(\max\{x,y\} = m \mid \min\{x,y\} = 2)$ for $m = 1, 2, 3, 4$.

Bayes' rule:

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

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

$$\Rightarrow P(A|B)P(B) = P(B|A)P(A) \quad \text{for } P(A)P(B) > 0.$$

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

Multiplication Rule

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1)P(A_2|A_1)\cdots P\left(A_n \Big| \bigcap_{i=1}^{n-1} A_i\right).$$

$$P(A_1 \cap A_2 \cap A_3) = P(A_1 \cap A_2)\,P(A_3|A_1 \cap A_2)$$

$$= P(A_1)\,P(A_2|A_1)\,P(A_3|A_1 \cap A_2)$$

Proof of the multiplication rule follows by mathematical induction.

Total Probability Theorem

Let $A_1, A_2, \ldots, A_n$ be mutually exclusive and collectively exhaustive events of the sample space (each possible outcome is included in exactly one of the events $A_1, A_2, \ldots, A_n$) and assume that $P(A_i) > 0 \ \forall i$. Then for any event B we have

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

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

Proof. $P(B) = P(B \cap \Omega)$

$$= P(B \cap (A_1 \cup A_2 \cup \dots \cup A_n))$$

$$= P\left(\bigcup_{i=1}^{n} (B \cap A_i)\right)$$

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

Bayes' Rule (Refined version)

Let $A_1, A_2, \dots, A_n$ be disjoint events that form a partition of the sample space and assume that $P(A_i) > 0 \ \forall i$. Then for any event B such that $P(B) > 0$ we have

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

$$= \frac{P(A_i)\,P(B|A_i)}{\sum_{j=1}^{n} P(A_j)\,P(B|A_j)}.$$

Example. (Exercise 3.4 in class)

Let $A = \{\text{an aircraft is present}\}$,

$B = \{\text{the radar generates an alarm}\}$.

We are given that

$P(A) = 0.05 \quad P(B|A) = 0.99 \quad P(B|A^c) = 0.1$.

$P(\text{aircraft is present} | \text{radar generates alarm})$

$= P(A|B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{P(B|A)P(A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}$

$= \dfrac{0.05 \times 0.99}{0.05 \times 0.99 + 0.95 \times 0.1}$.

Independence. Two events A & B are called independent events if $P(A \cap B) = P(A) \cdot P(B)$.

$P(A|B) = P(A)$ for $P(B) > 0$, or

$P(B|A) = P(B)$ for $P(A) > 0$.

Exercise 3.2. If A and B are independent events. Show that

(i) A^c and B are independent

(ii) A^c and B^c are independent.

Lecture 4

(17 August 2023)

Independence of several events

We say that the events $A_1, A_2, \ldots, A_n$ are independent if

$$P\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} P(A_i),$$

for every subset S of $\{1, 2, \ldots, n\}$.

For three events A, B and C:

$$P(A \cap B) = P(A)\,P(B) \longrightarrow (1)$$

$$P(B \cap C) = P(B)\,P(C) \longrightarrow (2)$$

$$P(A \cap C) = P(A)\,P(C) \longrightarrow (3)$$

$$P(A \cap B \cap C) = P(A)\,P(B)\,P(C) \longrightarrow (4)$$

(1)–(3) do not imply (4)

(4) does not imply (1)–(3)

Exercise 4.1. Consider two independent fair coin tosses, and the following events:

$H_1 = \{HT, HH\}$,

$H_2 = \{TH, HH\}$,

$D = \{TT, HH\}$.

Show that H_1, H_2 and D are pairwise independent but not independent.

Exercise 4.2. Consider two independent rolls of a fair six-sided die and the following events:

$A = \{1^{st}$ roll is 1, 2 or $3\}$

$B = \{1^{st}$ roll is 3, 4 or $5\}$

$C = \{$the sum of the two rolls is $9\}$.

Show that ④ holds but ①-③ do not hold.

Conditional Independence

Given an event C with $P(C) > 0$, the events A and B are called independent if

$$P(A \cap B \mid C) = P(A \mid C)\, P(B \mid C).$$

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

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

$$= P(B \mid C)\, P(A \mid B \cap C).$$

Assume $P(B \mid C) \neq 0$ the above definition can be written as

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

In words this states that if C is known to have occurred, an additional knowledge that B also occurred does not change the probability of A.

Interestingly independence of two events A and B does not imply conditional independency and vice versa.

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

$$\not\Leftrightarrow \quad P(A \cap B \mid C) = P(A \mid C)\, P(B \mid C)$$

in general.

Exercise 4.3. Consider two independent fair coin tosses, and the events same as in Exercise 4.1.

$H_1 = \{HT, HH\}$

$H_2 = \{TH, HH\}$

$D = \{HT, TH\}$.

$P(H_1 \cap H_2) = P(H_1)P(H_2)$ and $P(H_1 \cap H_2 | D) \neq P(H_1|D)P(H_2|D)$.

Exercise 4.4. There are two coins: a fair coin and a fake two-headed coin ($P(H)=1$). We choose one of the two coins at random, each being chosen with probability $\frac{1}{2}$ and toss it twice.

H_1, H_2 same as Exercise 4.3.

$F = \{$fair coin is selected$\}$.

$P(H_1 \cap H_2 | F) = P(H_1|F)P(H_2|F)$ and

$P(H_1 \cap H_2) \neq P(H_1)P(H_2)$.

Review of Counting.

Permutations: Given n distinct objects, and let $k \leq n$, we wish to count the number of different ways that we can pick k out of these n objects and arrange them in a sequence, i.e., the number of distinct k-object sequences.

$$ {}^{n}P_{k} = \frac{n!}{(n-k)!} = n(n-1)\cdots(n-k+1). $$

Combinations: Counting the number of k-element subsets of a given n-element set. Notice that forming a combination is different than forming a permutation, because in a combination there is no ordering of the selected elements.

$$ \binom{n}{k} = {}^{n}C_{k} = \frac{n!}{(n-k)!\,k!} $$

Partitions : Consider n and $n_1 \cdots n_r$ s.t.

$$n = n_1 + n_2 + \cdots + n_r .$$

No. of partitions of n distinct elements into r disjoint subsets with the ith subset containing exactly n_i elements

$$= \frac{n!}{n_1!\, n_2! \cdots n_r!}$$

Continuity of Probability

Notation. $\bigcup_{i=1}^{\infty} A_i \triangleq \lim_{n\to\infty} \bigcup_{i=1}^{n} A_i$,

$$\sum_{i=1}^{\infty} P(A_i) \triangleq \lim_{n\to\infty} \sum_{i=1}^{n} P(A_i)$$

Theorem.

$$\lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} A_i\right) = P\left(\lim_{n\to\infty} \bigcup_{i=1}^{n} A_i\right) .$$

proof. $B_1 = A_1$

$$B_i = A_i \setminus \bigcup_{j=1}^{i-1} A_j , \quad i = 2, 3, \ldots$$

(i) B_i's are disjoint (verify)

(ii) $\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$ — $n \in \mathbb{N}$ and

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i .$$

Let $C_n = \bigcup_{i=1}^{n} A_i$. $C_2 = B_1 \cup B_2$

(base case)

Assume $C_{n-1} = \bigcup_{i=1}^{n-1} B_i$.

$$C_n = C_{n-1} \cup A_n = C_{n-1} \cup (A_n \setminus C_{n-1})$$

$$= C_{n-1} \cup \left(A_n \setminus \bigcup_{i=1}^{n-1} A_i \right)$$

$$= C_{n-1} \cup B_n = \bigcup_{i=1}^{n} B_i .$$

$x \in \bigcup_{i=1}^{\infty} A_i \Rightarrow \exists\, n$ s.t. $x \in A_n \Rightarrow x \in \bigcup_{i=1}^{n} B_i$

$$\Rightarrow x \in \bigcup_{i=1}^{\infty} B_i$$

likewise $x \in \bigcup_{i=1}^{\infty} B_i \Rightarrow x \in \bigcup_{i=1}^{\infty} A_i$.

Now we have

$$\lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} A_i\right) = \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} B_n\right)$$

$$= \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i) \text{ (by additivity)}$$

$$= \sum_{i=1}^{\infty} P(B_i)$$

$$= P\left(\bigcup_{i=1}^{\infty} B_i\right) \text{ (by additivity)}$$

$$= P\left(\bigcup_{i=1}^{\infty} A_i\right)$$

$$= P\left(\lim_{n\to\infty} \bigcup_{i=1}^{n} A_i\right).$$

∎

Consequences.

1) Let $A_1, A_2, \ldots$ be a monotonically increasing sequence, i.e., $A_i \subseteq A_{i+1}$, $i \in \mathbb{N}$.

Then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} P(A_n)$.

2) Let $A_1, A_2, \ldots$ be a monotonically decreasing sequence, i.e. $A_{i+1} \subseteq A_i$ $i \in \mathbb{N}$.

Then

$$P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{n \to \infty} P(A_n).$$

This follows from the previous consequence by considering complements of the events and using Demorgan's laws.

Lecture 5

(19 August 2023)

Continuity of probability

$$\lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} A_i\right) = P\left(\lim_{n\to\infty} \bigcup_{i=1}^{n} A_i\right).$$

$$B_i = A_i \setminus \bigcup_{j=1}^{i-1} A_j, \quad i \in N.$$

B_i and $B_{i'}$ are disjoint for $i \neq i'$.

Let $i < i'$ without loss of generality.

Let $x \in B_i$.

$\Rightarrow x \in A_i$

$\Rightarrow x \notin A_{i'} \setminus \bigcup_{j=1}^{i'-1} A_j$ since $A_i \subseteq \bigcup_{j=1}^{i'-1} A_j$

$\Rightarrow x \notin B_{i'}$.

Let $x \in B_{i'} \Rightarrow x \in A_{i'} \setminus \bigcup_{j=1}^{i'-1} A_j$

$\Rightarrow x \notin A_i \Rightarrow x \notin B_i$

Therefore B_i and $B_{i'}$ are disjoint sets

$$\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i .$$

We prove this using mathematical induction.

Let $C_n = \bigcup_{i=1}^{n} A_i$ $\quad n = 1, 2, \ldots$

$C_2 = A_1 \cup A_2 = B_1 \cup B_2$ (base case)

Assume that $C_{n-1} = \bigcup_{i=1}^{n-1} B_i$.

Consider

$$C_n = C_{n-1} \cup A_n$$

$$= C_{n-1} \cup (A_n \setminus C_{n-1})$$

$$= C_{n-1} \cup \left(A_n \setminus \bigcup_{i=1}^{n-1} B_i\right)$$

$$= C_{n-1} \cup B_n = \bigcup_{i=1}^{n} B_n .$$

Also $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$.

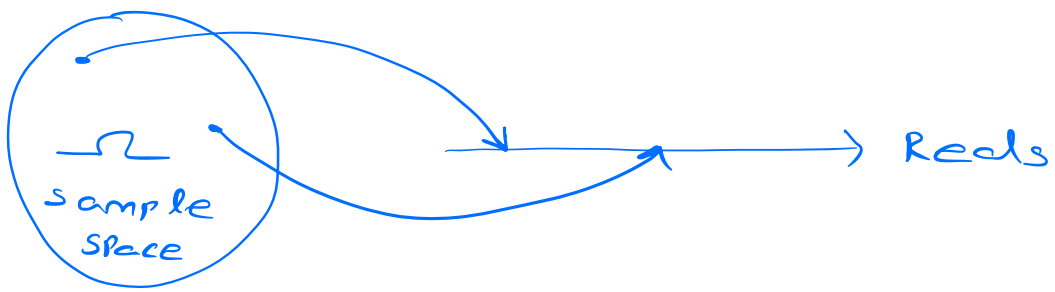
Let $x \in \bigcup_{i=1}^{\infty} A_i \Rightarrow \exists\, n \in \mathbb{N}$ s.t. $x \in A_n$

$\Rightarrow x \in \bigcup_{i=1}^{n} A_i \Rightarrow x \in \bigcup_{i=1}^{n} B_i \Rightarrow x \in \bigcup_{i=1}^{\infty} B_i$.

Similarly $x \in \bigcup_{i=1}^{\infty} B_i \Rightarrow x \in \bigcup_{i=1}^{\infty} A_i$.

Random Variables

In many probabilistic models the outcomes are not numerical (real values) but they may be associated with some numerical values of interest.



Formally a random variable is a real valued function of the experimental outcome.

$$X : \Omega \to R .$$

Coin toss $\Omega = \{H, T\}$

$$X(H) = 1, \quad X(T) = 0.$$

Roll a die twice

X = maximum of two rolls

Definition. A discrete random variable is a real-valued function of the outcome of the experiment that can take a finite or countably infinite number of values.

- A discrete random variable has an associated probability mass function (PMF), which gives the probability of each numerical value that the random variable can take.

Definition. Consider a probabilistic model with sample space Ω and probability law P. Let X be a random variable, $X : \Omega \to \mathbb{R}$. Probability mass of $x \in \mathbb{R}$ is defined as

$$P_X(x) = P(\{\omega \in \Omega : X(\omega) = x\}).$$

Notation. $\{X = x\} \triangleq \{\omega \in \Omega : X(\omega) = x\}$.

Example. Consider two independent tosses of a fair coin, and let X be no. of heads obtained.

$$P_X(x) = \begin{cases} 1/4 & \text{if } x = 0 \text{ or } x = 2 \\ 1/2 & \text{if } x = 1 \\ 0 & \text{o.w.} \end{cases}$$

Let $X \subseteq \mathbb{R}$ be the range of the function $X : \Omega \to \mathbb{R}$.

<u>Claim</u>. $\sum_{x \in X} p_X(x) = 1$.

Proof. $\sum_{x \in X} p_X(x) = \sum_{x \in X} P(X = x)$

$= \sum_{x \in X} P(\underbrace{\{\omega : X(\omega) = x\}}_{A_x})$

[A_x, $x \in X$ are disjoint events and form a partition of Ω]

$= P\left(\bigcup_{x \in X} A_x\right) = P(\Omega) = 1$.

(by additivity) (normalization)

- We denote the probability that X takes a value within a set $S \subseteq \mathbb{R}$ by

$$P(X \in S) \triangleq \sum_{x \in S} p_X(x).$$

<u>Example</u>. If X is the no. of heads obtained in two independent tosses of a fair coin

the probability of at least one head is

$$P(X>0) = P(X \in \{1,2\})$$

$$= \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

The Bernoulli Random variable

Consider the toss of a coin which comes up a head with probability p and a tail with a probability $1-p$.

$$X(H) = 1 \quad X(T) = 0.$$

$$P_X(1) = p \quad P_X(0) = 1-p.$$

This is a very important r.v. (random variable). In practice it is used to model generic probabilistic situations with just two outcomes.

By combining multiple Bernoulli r.v.'s we get Binomial random variable.

Binomial Random Variable

A coin is tossed n times (independently)

$P(H) = p \quad P(T) = 1-p.$

Let X be the no. of heads in the n-toss sequence $\longrightarrow$ Binomial r.v.

$$p_X(k) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, \ldots, n.$$

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

Lecture 6

(21 August 2023)

Geometric Random variable

Suppose that we repeatedly and independently toss a coin until a head comes up for the first time. Let $P(H) = p$, $0 < p < 1$.

Geometric r.v. is the no. of tosses needed for a head to come up for the first time.

$$P_X(k) = (1-p)^{k-1} p \quad k = 1, 2, \ldots,$$

Since $(1-p)^{k-1} p$ is the probability of the sequence consisting of $k-1$ successive tails followed by a head.

This is a valid PMF because

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

$$= p \cdot \sum_{k=0}^{\infty} (1-p)^k$$

$$= p \cdot \frac{1}{1-(1-p)} = 1.$$

Poisson Random variable

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

where λ is a positive parameter characterizing the PMF. This is a valid PMF because

$$\sum_{k=0}^{\infty} e^{-\lambda} \cdot \frac{\lambda^k}{k!} = e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

Theorem. Consider a binomial distribution with parameters n and p. As $n \to \infty$ and $p = \lambda/n$, while keeping λ constant, we have

$$\lim_{n \to \infty} \frac{n!}{(n-k)!\, k!} p^k (1-p)^{n-k} = e^{-\lambda} \frac{\lambda^k}{k!}.$$

Proof.

$$\frac{n!}{(n-k)!\, k!} p^k (1-p)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \underbrace{\frac{n(n-1)\cdots(n-k+1)}{n^k}}_{\to 1} \cdot \frac{\lambda^k}{k!} \underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\to 1} \left(1-\frac{\lambda}{n}\right)^n$$

$$\lim_{n\to\infty} (1-\lambda/n)^n \longrightarrow e^{-\lambda}.$$

This gives $\dfrac{e^{-\lambda}\cdot\lambda^k}{k!}$.

When n is very large and p is very small Poisson PMF is a good approximation to binomial PMF.

Example. $n=100$ $p=0,01$. The probability of 5 successes in 100 trials

$$= \binom{100}{5} (0,01)^5 (0,09)^{95} = 0,00290.$$

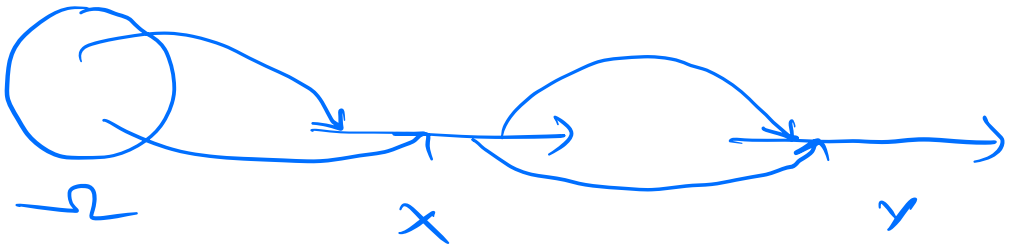
Using the Poisson PMF as an approximation with $\lambda = np = 1$, the probability equals

$$e^{-1}\frac{1}{5!} = 0,00306.$$

Functions of Random Variables

Let x be a random variable, i.e., $x : \Omega \to R$, and y be a

function of x, i.e., $y = g(x)$, $g: R \to R$.



Y is also a random variable,

$$P_Y(y) = \sum_{x: g(x)=y} P_X(x),$$

→ Exercise 6.1: Prove this

Example. Let $Y = |X|$, and

$$P_X(x) = \begin{cases} 1/9, & \text{if } x \text{ is an integer in } [-4, 4] \\ 0, & 0, \omega. \end{cases}$$

Find the PMF of Y.

Proof. $X = [-4 : 4]$,

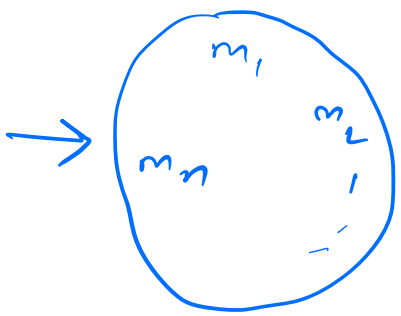
$Y = [0 : 4]$.

$P_Y(0) = 1/9$, $P_Y(i) = \frac{2}{9}$, $i \in [1 : 4]$,

$P_Y(y) = 0$, $0, \omega$.

Expectation. The PMF of X provides us with several numbers, the probabilities of all the possible values of X. It is often desirable, however, to summarize this information in a single representative number. This is accomplished by the expectation of X, which is a weighted average of the possible values of X.

Motivation. wheel of fortune



m_i comes up with probability p_i

$i \in [1:n]$.

Suppose you spin the wheel k times,

Let k_i be the no. of times that the outcome is m_i. $\quad k = \sum_{i=1}^{n} k_i$.

Total reward $= \sum_{i=1}^{n} m_i k_i$.

Total reward per spin $= \sum_{i=1}^{n} m_i k_i / k$.

If we interpret probabilities as relative

frequencies $\frac{k_i}{k} \to p_i$ as $k \to \infty$.

$$\frac{\sum_{i=1}^{n} m_i k_i}{k} \longrightarrow \sum_{i=1}^{n} m_i p_i .$$

Expectation (or Expected value or mean)

Expectation of a RV x with PMF p_x is defined by

$$E[x] = \sum_{x \in \mathcal{X}} x\, p_x(x),$$

where $\mathcal{X}$ is the set of all possible values that x can take.

Mean of a Bernoulli RV with $p_x(1) = p$ is

$$E[x] = 1 \cdot p + 0 \cdot (1-p) = p.$$

Consider a RV x with $p_x(2^k) = \frac{1}{2^k}$, $k = 1, 2, \ldots$

$$E[x] = \sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2^k} = \infty$$

Thus, expectation may not be always well defined. However there exists RVs

taking infinite number of values with finite mean.

Exercise 6.2. Find the expectation of a geometric RV with probability of heads equal to p, where $0<p<1$.

Lecture 7

(24 August 2023)

Variance. $\mathrm{Var}(X) = E[(X - E[X])^2]$.

Since $(X - E[X])^2$ can only take nonnegative values, the variance is always nonnegative.

The variance provides a measure of dispersion of X around its mean.

$\sigma_X = \sqrt{\mathrm{Var}(X)}$ is called the standard deviation.

- Can variance be zero?

Yes!

We will get back to this soon.

P.T.O

Example. $p_X(x) = \begin{cases} 1/9, & x \in [-4:4] \\ 0, & \text{o.w.} \end{cases}$

Compute the variance $\text{Var}(X) = E[(X - EX)^2]$.

$E[X] = 0$.

Let $Y = (X - E[X])^2 \in \{0, 1, 4, 9, 16\}$.

$p_Y(0) = 1/9$, $p_Y(i^2) = \frac{2}{9}$, $i \in [1:4]$.

$$\text{Var}(X) = E[Y] = 0 \cdot \frac{1}{9} + 1 \cdot \frac{2}{9} + 4 \cdot \frac{2}{9} + 9 \cdot \frac{2}{9} + 16 \cdot \frac{2}{9}$$

$$= \frac{60}{9}.$$

Expected value of Functions of Random variables

Theorem.

Let X be a RV and $g(X)$ be a function of X. Then the expected value of $g(X)$ is given by

$$E[g(X)] = \sum_{x \in \mathcal{X}} x \, p_X(x).$$

Proof. Let $Y = g(X)$, $\mathcal{Y} = \{y \in \mathbb{R} : \exists x \text{ s.t. } g(x) = y\}$.

$$E[g(x)] = E[Y] = \sum_{y \in Y} y P_Y(y)$$

$$= \sum_{y \in Y} y \sum_{x : g(x) = y} P_X(x)$$

$$= \sum_{y \in Y} \sum_{x : g(x) = y} y P_X(x)$$

$$= \sum_{y \in Y} \sum_{x : g(x) = y} g(x) P_X(x)$$

$$= \sum_{x \in X} g(x) P_X(x).$$

Using this

$$\text{Var}(X) = E[(X - E[X])^2] = \sum_x (x - E[X])^2 P_X(x).$$

For the example above, $\text{Var}(X) = \sum_{x \in [-4:4]} (x-0)^2 P_X(x) = \frac{60}{9}$.

$\rightarrow \text{Var}(X) = 0 \implies x = E[X]$ for all x. X is a constant.

Properties of mean and variance.

- $Y = aX + b$

$$E[Y] = a E[X] + b$$

$$\text{Var}(Y) = E\left[(Y - E[Y])^2\right]$$

$$= E\left[\left(ax + b - aE[x] - b \right)^2 \right]$$

$$= E\left[a^2 (x - E[x])^2 \right] = a^2 \, Var(x)$$

Variance in Terms of moments

$$Var(x) = E[x^2] - (E[x])^2$$

Mean and Variance of some common RVs

Bernoulli RV:

$$P_x(1) = p = 1 - P_x(0)$$

$$E[x] = p, \quad E[x^2] = p, \quad Var(x) = p - p^2 = p(1-p)$$

Discrete uniform random variable:

$$P_x(k) = \frac{1}{b-a+1},$$

$$k \in [a : b]$$

$$E[x] = \frac{a+b}{2}$$

Let $y \sim$ uniform $[1 : n]$; $y = x - a + 1 \Rightarrow n = b - a + 1$

$$Var(y) = \frac{n^2 - 1}{12}, \quad E[y^2] = \sum_{k=1}^{n} k^2 / n = \frac{1}{6}(n+1)(2n+1)$$

General case: $Var(Y) = \frac{(b-a+1)^2 - 1}{12}$, for $n = b-a+1$,

since $Var(X) = Var(X-a+1) = Var(Y)$,

$X-a+1 \in [1:n]$ for $n = b-a+1$,

Binomial random variable:

$$P_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in [0:n].$$

$$E[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

$$k\binom{n}{k} = k \cdot \frac{n!}{(n-k)!\,k!} = \frac{n!}{(n-k)!\,(k-1)!}$$

$$= n \binom{n-1}{k-1}.$$

$$E[X] = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k}$$

$n' = n-1 \quad k' = k-1$

$$= np \sum_{k'=0}^{n'} \binom{n'}{k'} p^{k'} (1-p)^{n'-k'}$$

$$= np\,(p+1-p)^{n'} = np.$$

$$E[X^2] = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k}$$

$$= np \sum_{k=1}^{n} k \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k}$$

$$\left[n' = n-1 \quad k' = k-1 \right]$$

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

$$= np\left(1 + (n-1)p\right)$$

$$\mathrm{Var}(X) = E[X^2] - E[X]^2$$

$$= np(1+(n-1)p) - n^2p^2$$

$$= np(1 + np - p - np) = np(1-p)$$

Exercise. Find the mean and variance of geometric random variable with probability of success $0 < p < 1$.

$$p_X(k) = (1-p)^{k-1} p \quad k \in \mathbb{N}.$$

Exercise. Find the mean and variance of Poisson RV with PMF $p_X(k) = e^{-\lambda} \lambda^k / k!$, $k \in 0, 1, 2, \ldots$

Lecture 8

(31 August 2023)

Joint PMFs of Multiple Random variables

Consider two RVs associated with the same experiment. The probabilities of the values that x and y can take are captured by the joint PMF of x and y denoted P_{xy}.

$$X : \Omega \to R,$$
$$Y : \Omega \to R.$$

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

$$=: P(X=x, Y=y).$$

$$P((X,Y) \in A) = \sum_{(x,y) \in A} P_{XY}(x,y).$$

Let $R_{XY} = \{(x,y) : P_{XY}(x,y) > 0\}$

$$R_X = \{x : P_X(x) > 0\}$$

$$R_Y = \{y : P_Y(y) > 0\}.$$

In fact, we can calculate the PMFs of X and Y by using the formulas

$$P_X(x) = \sum_{y \in R_Y} P_{XY}(x,y) \qquad P_Y(y) = \sum_{x \in R_X} P_{XY}(x,y).$$

$$P_X(x) = P(X = x)$$

$$= P(X = x \cap \Omega)$$

$$= P\left(X = x \cap \bigcup_{y \in R_Y} Y = y\right)$$

$$= \sum_{y \in R_Y} P(X = x \cap Y = y)$$

$$= \sum_{y \in R_Y} P_{XY}(x,y).$$

Also $$\sum_{(x,y) \in R_{XY}} P_{XY}(x,y) = \sum_{(x,y) \in R_{XY}} P(X = x, Y = y)$$

$$= \sum_{(x,y) \in R_{XY}} P(A_{xy})$$

$$= P\left(\bigcup_{(x,y) \in R_{XY}} A_{xy}\right) = P(\Omega) = 1.$$

$A_{xy}, (x,y) \in R_{XY}$ form a partition of Ω.

Functions of multiple Random variables

Let $Z = g(X,Y)$ be a function of RVs X and Y.

$$P_Z(z) = \sum_{(x,y):\, g(x,y)=z} P_{XY}(x,y).$$

Exercise. Show that

$$E[g(X,Y)] = \sum_{x,y} g(x,y)\, P_{XY}(x,y).$$

Linearity of Expectation

$g(X,Y) = X+Y$

$$E[X+Y] = \sum_{x,y} (x+y)\, P_{XY}(x,y)$$

$$= \sum_x x P_X(x) + \sum_y y P_Y(y) = E[X] + E[Y].$$

More than Two Random Variables

$$P_{XYZ}(x,y,z) = P(X=x, Y=y, Z=z).$$

$$P_X(x) = \sum_{(y,z)\in R_{YZ}} P_{XYZ}(x,y,z),$$

$$P_{YZ}(y,z) = \sum_{x\in R_X} P_{XYZ}(x,y,z),$$

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

Example. mean of the Binomial distribution

$$X = X_1 + X_2 + \cdots + X_n \quad \text{(Binomial)}$$

$$P(X_i = 1) = p = 1 - P(X_i = 0) \quad \text{(Bernoulli)}$$

$$E[X] = np.$$

Conditioning a RV on an Event

The conditional PMF of a RV X conditioned on a particular event A with $P(A) > 0$ is defined by

$$p_{X|A}(x) = P(X = x \mid A)$$

$$= P(\{X = x\} \cap A) / P(A).$$

$$\sum_x p_{X|A}(x) = \sum_x \frac{P(\{X = x\} \cap A)}{P(A)} = \frac{P(A)}{P(A)} = 1.$$

Example. Let X be the roll of a fair six-sided die and let A be the event that the roll is an even number.

$$p_{X|A}(k) = \frac{P(X = k \text{ and } X \text{ is even})}{P(\text{roll is even})}$$

$$= \begin{cases} 1/3, & \text{if } k = 2, 4, 6 \\ 0, & \text{otherwise.} \end{cases}$$

$p_X(x) = \sum_{i=1}^{n} P(A_i) p_{X|A_i}(x)$ if $A_1, \ldots, A_n$ form a partition of Ω.

Conditioning a RV on another RV

Consider two RVs X and Y associated with the same experiment.

If we know that the value of Y is some particular y with $p_Y(y) > 0$, this provides partial knowledge about the value of X.

This knowledge is captured by the conditional PMF $p_{X|Y}$ of X given Y,

$$p_{X|Y}(x|y) = P(X = x \mid Y = y)$$

$$= \frac{P(X = x,\ Y = y)}{P(Y = y)} = \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

$$\sum_x p_{X|Y}(x|y) = 1.$$

$$p_{X,Y}(x,y) = p_X(x)\, p_{Y|X}(y|x)$$

$$= p_Y(y)\, p_{X|Y}(x|y)$$

Example. In each lecture a Professor is asked 0, 1, or 2 questions with equal probability $\frac{1}{3}$. He answers each question incorrectly with probability $\frac{1}{4}$ independent of other questions. Let x and y be the no. of questions the professor is asked and the no. of questions he answers wrong in a given lecture, respectively. Find $P_{xy}(x,y)$.

$$P_{xy}(0,k) = \begin{cases} \frac{1}{3} \cdot 1 = \frac{16}{48} & \text{if } k=0 \\ 0 & \text{if } k=1,2. \end{cases}$$

$$P_{y|x}(0|1) = \frac{3}{4}, \quad P_{y|x}(1|1) = \frac{1}{4}, \quad P_{y|x}(2|1) = 0,$$

$$P_{y|x}(0|2) = \frac{9}{16}, \quad P_{y|x}(1|2) = \frac{3}{8}, \quad P_{y|x}(2|2) = \frac{1}{16}.$$

P. T. O

Conditional Expectation

For an event A with $P(A) > 0$ the condition expectation $E[X|A]$ is defined by

$$E[X|A] = \sum_x x \, p_{X|A}(x).$$

$$E[g(X)|A] = \sum_x g(x) \, p_{X|A}(x).$$

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

Total Expectation Theorem:

If $A_1, \dots, A_n$ be disjoint events that form a partition of the sample space Ω, with $P(A_i) > 0$ for all i then

$$E[X] = \sum_{i=1}^{n} P(A_i) \, E[X|A_i].$$

Proof. $E[X] = \sum_x x \, p_X(x)$

$$= \sum_x x \left[\sum_{i=1}^{n} P(A_i) \, p_{X|A_i}(x) \right]$$

$$= \sum_{i=1}^{n} P(A_i) \left[\sum_x x \, P_{X|A_i}(x) \right]$$

$$= \sum_{i=1}^{n} P(A_i) \, E[X|A_i],$$

Also

$$E[X] = \sum_y P_Y(y) \, E[X|Y=y].$$

Similarly

$$E[X|B] = \sum_{i=1}^{n} P(A_i|B) \, E[X|A_i \cap B],$$

Lecture 9

(4 September 2023)

Conditional Expectation as a RV

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

Let $g(y) = E[X|Y=y]$.

Consider the RV $g(Y)$.

Define $E[X|Y] := g(Y)$.

Law of Iterated Expectations:

Let X and Y be two random variables distributed according to a joint PMF P_{XY}.

Then $E[E[X|Y]] = E[X]$.

Proof. $E[E[X|Y]] = E[g(Y)]$

$$= \sum_y g(y) P_Y(y)$$

$$= \sum_y E[X|Y=y] P_Y(y)$$

$$= \sum_y \sum_x x \, p_{X|Y}(x|y) \, p_Y(y)$$

$$= \sum_x x \sum_y p_{XY}(x,y) = \sum_x x \, p_X(x) = E[X].$$

Minimum Mean Square Error Estimator

$(X,Y) \sim p_{XY}$. Given an observation y we need to estimate x.

Let $\hat{x} = f(y)$.

Theorem. $f(y) = g(y) = E[X|Y=y]$ minimizes the expected squared error

$$E[(X-\hat{x})^2] = E[(X-f(Y))^2].$$

proof. $E[(X-f(Y))^2]$

$$= \sum_y E[(X-f(y))^2 | Y=y] \, p_Y(y)$$

$$= \sum_y E[X^2 + f(y)^2 - 2f(y)X \mid Y=y] \, p_Y(y)$$

$$= \sum_y \left(f(y)^2 - 2f(y)E[X|Y=y] + E[X^2|Y=y] \right) p_Y(y)$$

Achieves the minimum at $f(y) = E[X|Y=y]$.

Conditional Variance

$$\text{Var}(X|Y=y) = E[X^2|Y=y] - E[X|Y=y]^2$$

$\text{Var}(X|Y)$ is defined as a function of RV, Y which takes a value $\text{var}(X|Y=y)$ when $Y=y$.

Exercise. Prove that

$$\text{Var}(X) = E[\text{var}(X|Y)] + \text{Var}(E[X|Y]).$$

Independence

- RV X is independent of the event A if $P(X=x \text{ and } A) = p_X(x) P(A)$, for all x,

(or) $p_{X|A}(x) = p_X(x)$, for all x if $P(A) > 0$.

Independence of RVs.

Two RVs X and Y are independent if the events $\{X=x\}$ and $\{Y=y\}$ are independent for all x, y, i.e.,

$$p_{XY}(x,y) = p_X(x)\, p_Y(y).$$

Conditional indep. $p_{XY|A}(x,y) = p_{X|A}(x)\, p_{Y|A}(y)$.

Example.

p_{XY}

x \ y	1	2	3	4
1	0	$\frac{1}{20}$	0	0
2	0	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{1}{20}$
3	$\frac{2}{20}$	$\frac{4}{20}$	$\frac{1}{20}$	$\frac{2}{20}$
4	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	0

Are x and y independent?

$$p_{XY}(1,1) = 0 \neq p_X(1)\,p_Y(1) > 0.$$

Are x and y independent conditioned on $A = \{X \geq 3,\ Y \leq 2\}$?

$$p_{XY|A}(3,1) = \frac{2}{9}, \quad p_{XY|A}(3,2) = 4/9,$$

$$p_{XY|A}(4,1) = \frac{1}{9}, \quad p_{XY|A}(4,2) = \frac{2}{9},$$

$$p_{X|A}(3) = \frac{2}{3} = 1 - p_{X|A}(4),$$

$$p_{Y|A}(1) = \frac{1}{3} = 1 - p_{Y|A}(2).$$

So x and y are not independent.

But x and y are independent conditioned on A.

Example. $X, Y \in \{0, 1\}$. P_{XY} is joint PMF.

Suppose $P_{XY}(0,1) = P_X(0) P_Y(1)$. Are X and Y independent?

$$P_{XY}(1,1) = P_Y(1) - P_{XY}(0,1)$$

$$= P_Y(1) - P_X(0) P_Y(1)$$

$$= P_Y(1) P_X(1).$$

Similarly $P_{XY}(x,y) = P_X(x) P_Y(y)$ for all $x, y \in \{0,1\}$.

Theorem. If X and Y are independent random variables, then $E[XY] = E[X] E[Y]$.

Proof.

$$E[XY] = \sum_{x,y} xy \, P_{XY}(x,y)$$

$$= \sum_{x,y} xy \, P_X(x) P_Y(y)$$

$$= \left(\sum_x x P_X(x) \right) \left(\sum_y y P_Y(y) \right)$$

$$= E[X] E[Y].$$

Exercise. If X and Y are independent

(i) Show that $g(X)$ and $h(Y)$ are independent.

(ii) Prove that $E[g(X) h(Y)] = E[g(X)] E[h(Y)]$.

Indicator Random Variables

For $A \subseteq \Omega$, $1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$.

Exercise

Two events A and B are independent if and only if 1_A and 1_B are independent RVs.

Independence of several Random Variables

Three RVs x, y and z are said to be independent if

$$P_{xyz}(x,y,z) = P_x(x)\, P_y(y)\, P_z(z) \quad \text{for all } x, y, z.$$

Independence of n RVs:

$$P_{x_1 x_2 \cdots x_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} P_{x_i}(x_i).$$

Variance of sum of Independent RVs

$$\text{Var}(X+Y)$$

$$= \text{Var}\left(\underbrace{X-E[X]}_{\tilde{X}} + \underbrace{Y-E[Y]}_{\tilde{Y}} \right)$$

$$= \text{Var}(\tilde{X}+\tilde{Y})$$

$$= E[(\tilde{X}+\tilde{Y})^2]$$

$$= E[\tilde{X}^2 + \tilde{Y}^2 + 2\tilde{X}\tilde{Y}]$$

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

where $\text{Cov}(X,Y) = E[(X-E[X])(Y-E[Y])]$.

If X and Y are independent, $\text{Cov}(X,Y)=0$.

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y).$$

- If $X_1, X_2, \ldots, X_n$ are independent RVs

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i).$$

- Consider n independent coin tosses with probability of heads p.

Let $X_i = 1\{ i^{th} \text{ coin toss is heads} \}$,

$i \in [1:n]$.

$$\sum_{i=1}^{n} X_i = \text{no. of heads}$$

$$\text{Var}\left(\sum_{i=1}^{n} X_i \right) = np(1-p).$$

Lecture 10

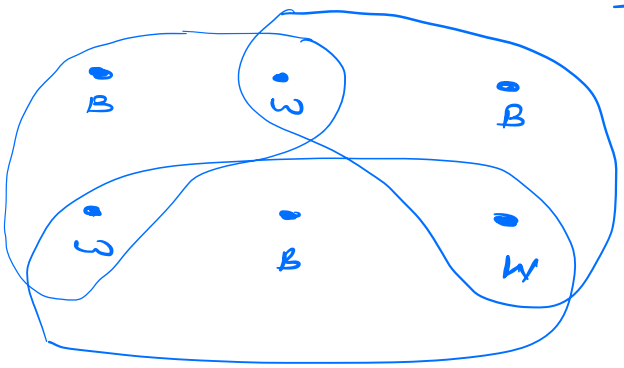
(7 September 2023)

Some Applications

Combinatorics and Graph Theory

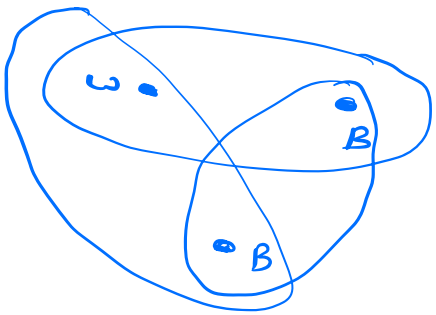
2-Coloring: [Existence Proof using union bound]

Let S be a set of some elements and $T_1, T_2, \ldots, T_m \subseteq S$ be subsets s.t. $|T_i| = \ell$ for $i \in [1:m]$. Can we 2-color S (meaning assign each element of S a color) such that each T_i has elements of both colors (i.e. not monochromatic)?



$\ell = 3 \quad m = 3$

2-coloring exists



$\ell = 2 \quad m = 3$

No 2-coloring

Theorem. Given S and $T_1, T_2, \ldots, T_m \subseteq S$ s.t. $|T_i| = l$ for $i \in [1; m]$, there exists a valid 2-coloring of S such that no T_i is monochromatic if $m < 2^{l-1}$.

For the examples above

$$l = 3,\ m = 3 \Rightarrow 3 < 2^{3-1} = 4$$

$$l = 2,\ m = 3 \Rightarrow 3 \not< 2^{2-1} = 3$$

Let $S = \{x_1, x_2, \ldots, x_n\}$.

Randomly color each element of x black or white independently and identically distributed each with probability $1/2$.

Let E_i be the event that T_i is monochromatic.

$$P(E_i) = \frac{1}{2^l} + \frac{1}{2^l} = \frac{1}{2^{l-1}}.$$

$$P\left(\bigcup_{i=1}^{m} E_i\right) \leq \sum_{i=1}^{m} P(E_i) = \frac{m}{2^{l-1}} < 1.$$

$$P(\text{each } T_i \text{ is not monochromatic})$$

$$= 1 - P\left(\bigcup_{i=1}^{\hat{}} E_i\right)$$

$$\geq 1 - \frac{m}{2^{l-1}} > 0.$$

Because we have a non-zero probability this implies that there exists a 2-coloring of S that gives all m valid non-monochromatic sets T_i.

Let $A = \{\omega \in \Omega$ satisfying some property$\}$

$P(A) > 0 \Rightarrow \exists\, \omega \in \Omega$ satisfying that property.

$\Rightarrow$

This is called the Probabilistic method.

Bipartite Subgraphs: [Existence proof using Linearity of Expectation]

Theorem, Let G be a simple graph with vertex set $[1:n]$ and m edges. Then G contains a bipartite subgraph with more than $m/2$ edges.

Proof. Let us split the vertices of G into two disjoint nonempty sets A and B.
($2^{n-1}-1$ such partitions)

Then A and B span a bipartite subgraph H of G (we remove the edges within A and within B).

$$\Omega = \{ \text{all partitions } (A, B) \}$$

$$|\Omega| = 2^{n-1} - 1.$$

Let $X(H) =$ no. of edges in H $H \in \Omega$.

Number the edges from 1 through m and let

$$X_i(A,B) = \begin{cases} 1 & \text{if the edge } i \text{ has one vertex in } A \text{ and one in } B \\ 0 & \text{otherwise.} \end{cases}$$

$$P(X_i = 1) = \frac{2^{n-2}}{2^{n-1}-1},$$ because we can get a subdivision of $[n]$ leading to $X_i = 1$ by first putting the two endpoints of the edge i to different subsets, then splitting

the remaining $(n-2)$-element vertex set in any of 2^{n-2} ways.

$$E[X_i] = 1 \cdot P_{X_i}(1) + 0 \cdot P_{X_i}(0)$$

$$= \frac{2^{n-2}}{2^{n-1}-1} > \frac{1}{2}.$$

$$X = \sum_{i=1}^{m} X_i \Rightarrow E[X] = \sum_{i=1}^{m} E[X_i] > \frac{m}{2}.$$

$$\Rightarrow \exists\, x \text{ s.t. } x > \frac{m}{2}.$$

$$\Rightarrow \exists\, H \in \Omega \text{ s.t. } X(H) = x > \frac{m}{2}.$$

- Minimum Mean Squared Error (MMSE)

The MMSE estimate of the random variable X, given that we have observed Y is given by $\hat{x}(y) = E[X|Y]$.

In particular, the estimate function $\hat{x}(y) = E[X|Y=y]$ achieves the minimum in

$$\min_{f(\cdot)} E[(X - f(Y))^2].$$

Proved in the last lecture.

Entropy (Uncertainity)

Consider a RV X_1 with PMF

$$P_{X_1}(0) = \tfrac{1}{2} = P_{X_1}(1).$$

Consider another r.v. X_2 with PMF

$$P_{X_2}(0) = 0.9 \quad P_{X_2}(1) = 0.1.$$

Which random variable's realization is hard to guess X_1 or X_2?

It appears that the uncertainity in X_1 is higher than that of X_2. This uncertainity is exactly captured by entropy.

For a RV, X with $P_X(1) = p = 1 - P_X(0)$

Entropy is defined as

$$H(X) = -P_X(1)\log P_X(1) - P_X(0)\log P_X(0)$$

$$= -p\log p - (1-p)\log(1-p),$$

$$=: h(p).$$

Is $H(X_2) \leq H(X_1)$? Yes!

$$H(X_1) = h(p) = -p\log p - (1-p)\log(1-p)$$

$$h'(p) = -p \cdot \frac{1}{p} - \log p + \frac{1-p}{1-p} + \log(1-p) = 0$$

$$\Rightarrow p = 1-p \Rightarrow p = \frac{1}{2}$$

$$h''(p) = \frac{-1}{p} - \frac{1}{1-p} < 0$$

$h(1/2) = 1$ is the maximum value of $h(p)$ according to the intuition X_1 has more uncertainity.

- Entropy is the fundamental quantity in information theory.

Lecture 11

(11 September 2023)

Recall that a random variable is a function from the sample space of a random experiment to the real numbers.

$$X : \Omega \to \mathbb{R}.$$

Let Ω be an uncountable set and P be the associated probability law.

For a set $B \subseteq \mathbb{R}$, $\{X \in B\} = \{\omega : X(\omega) \in B\}$,

i.e., $P(X \in B) = P(\{\omega : X(\omega) \in B\})$.

Continuous Random variable

A RV X is called continuous if there is a nonnegative function f_X, called the probability density function (PDF) of X such that

$$P(X \in B) = \int_B f_X(x)\,dx, \text{ for every } B \subseteq \mathbb{R}.$$

In particular, the probability that the value of x falls within an interval is

$$P(a \le x \le b) = \int_a^b f_x(x)\,dx.$$

(area under the graph of the PDF f_x)

$$P(x = a) = \int_a^a f_x(x)\,dx = 0$$

For this reason, including or excluding the endpoints of an interval has no effect on its probability:

$$P(a \le x \le b) = P(a < x \le b) = P(a \le x < b) = P(a < x < b).$$

- $f_x(x) \ge 0 \quad \forall x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f_x(x)\,dx = 1.$

This means that the entire area under the graph of the PDF must be equal to 1.

To interpret the PDF, note that for an interval $[x, x+\delta]$ with very small δ, we have

$$P(x \in [x, x+\delta]) = \int_x^{x+\delta} f_x(t)\,dt \approx f_x(x)\,\delta$$

$$f_x(x) \approx \frac{P(x \in [x, x+\delta])}{\delta}.$$

So we can view $f_x(x)$ as the 'probability mass per unit length' near x. It is important to realize that even though a PDF is used to calculate event probabilities $f_x(x)$ is not the probability of any particular event. In particular, it is not restricted to be less than or equal to 1.

Example. Uniform RV.

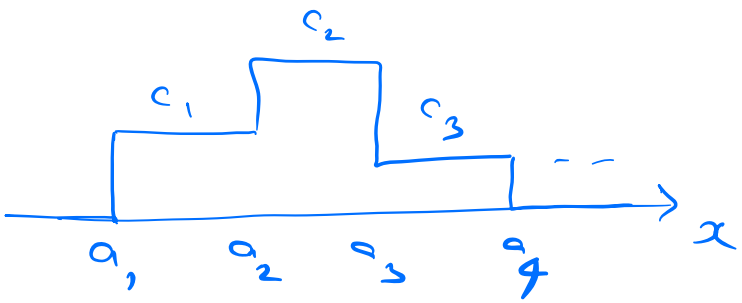
Consider a PDF

$$f_x(x) = \begin{cases} c & \text{if } a \le x \le b \\ 0 & \text{o.w.} \end{cases}$$

$$\int_a^b f_x(x)\,dx = 1 \Rightarrow c = 1/(b-a).$$

Example. Piecewise constant PDF.

$$f_X(x) = \begin{cases} c_i, & \text{if } a_i \le x \le a_{i+1}, \quad i = 1, 2, \dots, n-1 \\ 0, & \text{o.w.} \end{cases}$$



$$\int_{-\infty}^{\infty} f_X(x) = 1 \implies \sum_{i=1}^{n-1} c_i (a_{i+1} - a_i) = 1.$$

— A PDF can take arbitrarily large values, For example, uniform RV on [a,b] with $b - a < 1$.

Another example – $f_X(x) = \begin{cases} \frac{1}{2\sqrt{x}}, & \text{if } 0 < x \le 1 \\ 0, & \text{o.w.} \end{cases}$

Even though $f_X(x)$ becomes infinitely large as x approaches zero, this is still a valid PDF because

$$\int_{-\infty}^{\infty} f_X(x) = \sqrt{x}\Big|_0^1 = 1.$$

Expectation.

The expected value of a continuous RV X is defined by

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

- A function of continuous RV is also a RV. It can be continuous or discrete.

$Y = g(X)$.

$Y = g(X) = X$ Continuous RV

$$g(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}, \quad Y = g(X) \text{ is discrete.}$$

Theorem (Expected value rule for functions of RVs)

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

Proof. We will present a proof assuming that $g(X)$ is non-negative.

For a nonnegative continuous RV Y

$$E[Y] = \int_0^{\infty} P(Y > y)\, dy.$$

$$\int_0^\infty P(Y>y)\,dy = \int_0^\infty \int_y^\infty f_Y(t)\,dt\,dy$$

$$= \int_{t=0}^{\infty} \left(\int_{y=0}^{t} dy \right) f_Y(t)\,dt$$

$$= \int_{t=0}^{\infty} t\, f_Y(t)\,dt = E[Y],$$

Now for any function g s.t. $g(x) \geq 0$

$$E[g(x)] = \int_0^\infty P(g(x)>y)\,dy$$

$$= \int_0^\infty \int_{x:g(x)>y} f_X(x)\,dx\,dy$$

$$= \int_{x=-\infty}^{\infty} f_X(x) \int_{y=0}^{g(x)} dy\,dx$$

$$= \int_{x=-\infty}^{\infty} g(x) f_X(x)\,dx.$$

Exercise. Prove the above for a general

real-valued function g.

Hint: Show that $E[X] = \int_0^\infty P(X > x)\,dx - \int_0^\infty P(X < -x)\,dx$.

Variance. $\text{Var}(X) = E[(X - E[X])^2]$.

Properties

- $Y = aX + b$

$$E[Y] = aE[X] + b$$

$$\text{Var}(Y) = a^2 \text{Var}(X)$$

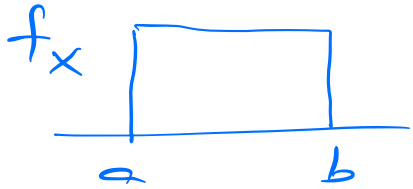
- $\text{Var}(X) = E[(X - E[X])^2]$

$$= \int_{x=-\infty}^{\infty} (x - E[X])^2 f_X(x)\,dx = E[X^2] - (E[X])^2.$$

Example. Mean and variance of the Uniform Random variable.

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{o.w.} \end{cases}$$

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_a^b \frac{1}{b-a} x\,dx = \frac{a+b}{2}.$$



$\Rightarrow E[X]$ = mid-point of $[a,b]$.

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

$$\text{Var}(X) = E[X^2] - (E[X])^2 = \frac{a^2+b^2+ab}{3} - \frac{(a+b)^2}{4}$$

$$= \frac{(b-a)^2}{12}.$$

Exponential Random variable

An exponential RV has a PDF of the form

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{o.w,} \end{cases}$$

where λ is a positive parameter characterizing the PDF.

This is a valid PDF because

$$\int_0^{\infty} \lambda e^{-\lambda x}\,dx = \lambda \left[\frac{e^{-\lambda x}}{-\lambda}\right]_0^{\infty} = 1,$$

- Exp. RV can model the amount of time until an incident of interest takes place,

The probability that x exceeds a certain value decreases exponentially.

Indeed for any $a \geq 0$, we have

$$P(X \geq a) = \int_a^\infty \lambda e^{-\lambda x} dx = e^{-\lambda a}.$$

$$E[X] = \int_0^\infty x \lambda e^{-\lambda x} dx$$

$$= x \lambda \frac{e^{-\lambda x}}{-\lambda} \Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx$$

$$= 0 + \frac{e^{-\lambda x}}{\lambda} \Big|_0^\infty = \frac{1}{\lambda}.$$

$$E[X^2] = \int_0^\infty x^2 \lambda e^{-\lambda x} dx$$

$$= x^2 (-e^{-\lambda x}) \Big|_0^\infty + \int_0^\infty 2x \cdot e^{-\lambda x} dx$$

$$= 0 + \frac{2}{\lambda} E[X] = \frac{2}{\lambda^2}.$$

$$\text{Var}(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2},$$

Lecture 12

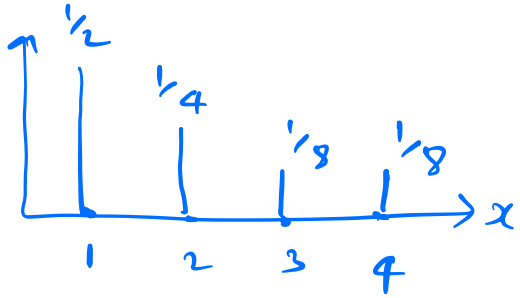
(14 September 2023)

Cumulative Distribution Functions

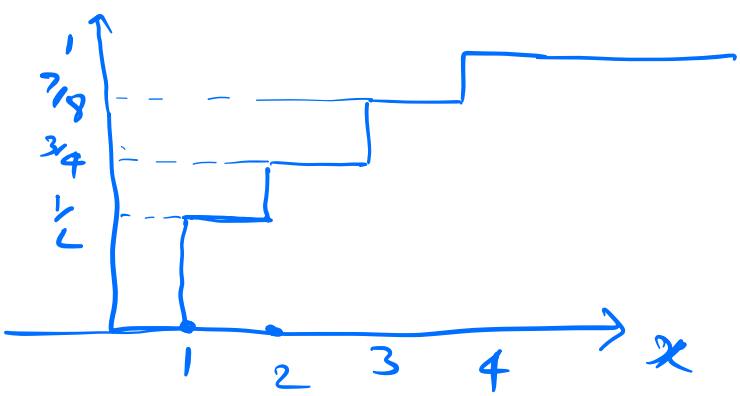
We would like to describe all kinds of RVs with a single mathematical concept. This is accomplished with CDF.

$$F_X(x) = \begin{cases} \sum_{k \le x} p_X(k) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{x} f_X(t)\,dt & \text{if } X \text{ is continuous.} \end{cases}$$

- Any random variable associated with a given probability model has a CDF regardless of whether it is discrete or continous. This is because $\{X \le x\}$ is always an event and therefor has a well-defined probability.

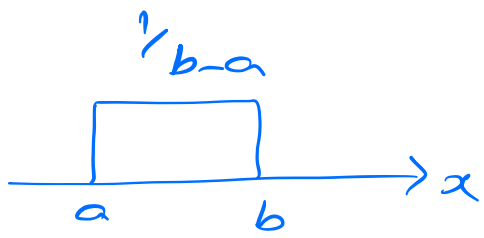


p_X

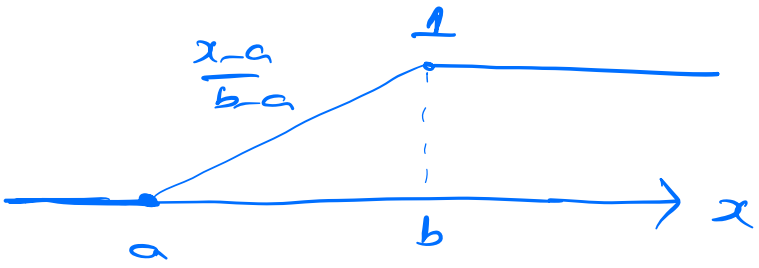


F_X

Discrete RV



f_x



Continuous RV

- CDF of discrete RV has jumps occuring at the values of positive probability mass. However F_x is continuous from the right.

Properties of a CDF

1) F_x is monotonically non-decreasing:

if $x \leq y$ then $F_x(x) \leq F_x(y)$.

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

$$\leq P(X \leq y) = F_y(y).$$

2) $\lim_{x \to -\infty} F_x(x) = 0$ $\quad \lim_{x \to \infty} F_x(x) = 1.$

$$A_n = \{ X \leq n \} \quad A_1 \subseteq A_2 \subseteq \cdots$$

$$\bigcup_{i=1}^{\infty} A_i = \Omega.$$

$$\lim_{x\to\infty} F_X(x) = \lim_{n\to\infty} P(X \le n)$$

Every sub-sequence of a monotonic sequence converges to the same limit

$$= P\left(\bigcup_{i=1}^{\infty} (X \le i)\right)$$

$$= P(\Omega) = 1.$$

Let $A_n = \{X \le -n\}$ $\quad A_1 \supseteq A_2 \supseteq \cdots$

$$\bigcap_{i=1}^{\infty} A_i = \phi$$

$$\lim_{x\to-\infty} F_X(x) = \lim_{x\to-\infty} P(X \le x)$$

$$= \lim_{n\to-\infty} P(X \le n)$$

$$= \lim_{n\to\infty} P(X \le -n)$$

$$= P\left(\bigcap_{n=1}^{\infty} (X \le -n)\right) = P(\phi) = 0.$$

3) F_X is right continuous.

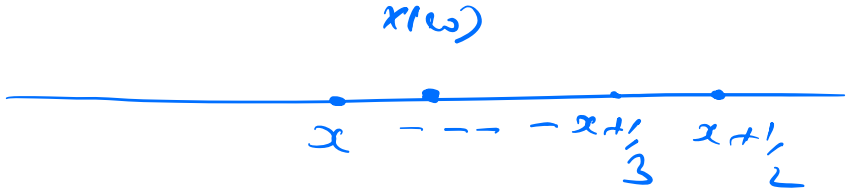
$$\lim_{\varepsilon\to 0+} F_X(x+\varepsilon) = F_X(x).$$

$$\lim_{\varepsilon \to 0^+} F_X(x+\varepsilon) = \lim_{n \to \infty} F_X(x+1/n)$$

$$= \lim_{n \to \infty} P(X \le x+1/n)$$

$A_n = \{X \le x+1/n\}$, $A_1 \supseteq A_2 \supseteq \cdots$

$$\bigcap_{i=1}^{\infty} A_i = \bigcap_{i=1}^{\infty} \{X \le x+1/i\} = \{X \le x\}$$



$x(\omega) \le x + 1/i \quad \forall i \in [1:\infty] \Rightarrow x(\omega) \le x$.

$$\lim_{\varepsilon \to 0^+} F_X(x+\varepsilon) = P\left(\bigcap_{i=1}^{\infty} A_i\right)$$

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

- If X is discrete, then F_X is a piecewise constant function of x.
- If X is continuous, then F_X is a continuous function of x.

$$\lim_{\varepsilon \to 0^-} F_X(x+\varepsilon) = \lim_{\varepsilon \to 0^-} P(X \le x+\varepsilon)$$

$$= \lim_{n\to\infty} P(x \le x - 1/n)$$

$A_1 \subseteq A_2 \subseteq \cdots$, $A_n = \{x \le x - 1/n\}$

$$= P\left(\bigcup_{i=1}^{\infty} A_i\right)$$

$$= P(x < x) \quad = P(x \le x) - \underbrace{P(x = x)}_{=0}$$

$$= F_x(x).$$

4) If x is discrete and takes integer values

$$F_x(k) = \sum_{i=-\infty}^{k} P_x(i),$$

$$P_x(k) = P(x \le k) - P(x \le k-1)$$

$$= F_x(k) - F_x(k-1), \quad \text{for all } k \in \mathbb{Z}.$$

If x is continuous

$$F_x(x) = \int_{-\infty}^{x} f_x(t)\,dt$$

$$f_x(x) = \frac{d F_x(x)}{dx}$$ for all x s.t. f is continuous at x.

[By Fundamental theorem of Calculus]

Lecture 13

(18 September 2023)

- Sometimes in order to calculate the PMF or PDF of a discrete or continuous r.v. respectively it is more convenient to first calculate the CDF.

Example. X_1, X_2 and X_3 are independent RVs with PMF $P_{X_i}(k) = \frac{1}{10}$ $k \in [1:10]$.

Find the PMF of $X = \max\{X_1, X_2, X_3\}$.

Proof. $F_X(k) = P(\max\{X_1, X_2, X_3\} \le k)$

$= P(X_1 \le k, X_2 \le k, X_3 \le k)$

$= P(X_1 \le k)\, P(X_2 \le k)\, P(X_3 \le k)$

since X_1, X_2 and X_3 are independent RVs

as $P(X_1 \le k, X_2 \le k, X_3 \le k)$

$= \sum_{a,b,c \le k} P_{X_1,X_2,X_3}(a,b,c) = \sum_{a,b,c \le k} P_{X_1}(a)\, P_{X_2}(b)\, P_{X_3}(c)$

$= P(X_1 \le k)\, P(X_2 \le k)\, P(X_3 \le k)$.

So $P(X \le k) = \left(\frac{k}{10}\right)^3$.

$P_X(k) = F_X(k) - F_X(k-1) = \left(\frac{k}{10}\right)^3 - \left(\frac{k-1}{10}\right)^3$

Q) Let A and B be events with $P(A) > 0$ and $P(B) > 0$, B suggests A if $P(A|B) > P(A)$ and B does not suggest A if $P(A|B) < P(A)$.

(a) Show that B suggests A if and only if A suggest B.

$$P(A|B) > P(A) \Rightarrow \frac{P(B|A)\,P(A)}{P(B)} > P(A)$$

$$\Rightarrow P(B|A) > P(B).$$

(b) If $P(B^c) > 0$ show that B suggests A if and only if B^c does not suggest A.

$P(A|B) > P(A)$

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

$$= \frac{P(A) - P(A \cap B)}{P(B^c)}$$

$$< P(A)\,P(B^c) / P(B^c)$$

Q) $X = \sum_{i=1}^{n} x_i$, x_i's are Bernoulli RVs (need not be independent. Show that

$$E[X^2] = \sum_{i=1}^{n} p_{x_i}(1)\, E[X \mid x_i = 1].$$

Solution,

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

$$= \sum_{i=1}^{n} E[x_i^2] + \sum_{\substack{i,j \\ i \neq j}} E[x_i x_j]$$

$$= \sum_{i=1}^{n} E[x_i] + \sum_{i \neq j} E[x_k x_j]$$

$$= \sum_{i=1}^{n} E\left[x_i \left(1 + \sum_{j \neq i} x_j\right)\right]$$

$$= \sum_{i=1}^{n} E\left[x_i \left(1 + (x - x_i)\right)\right]$$

$$= \sum_{i=1}^{n} E\left[x_i + x x_i - x_i^2\right]$$

$$= \sum_{i=1}^{n} E[x x_i] + \sum_{i=1}^{n} \underbrace{\left(E[x_i] - E[x_i^2]\right)}_{=0}$$

$$= \sum_{i=1}^{n} p_{x_i}(1)\, E[x \mid x_i = 1].$$

Q) Suppose that we flip a coin n times to obtain n random bits. Consider all $m = \binom{n}{2}$ pairs of these bits in some order. Let Y_i be the exclusive-or of the ith pair of bits and let $Y = \sum_{i=1}^{m} Y_i$ be the number of Y_i that equal 1.

(a) Show that each Y_i is 0 with probability $\frac{1}{2}$ and 1 with probability $\frac{1}{2}$.

(b) Show that Y_i's are not mutually independent.

(c) Show that $E[Y_i Y_j] = E[Y_i] E[Y_j] \quad \forall i, j$.

(d) Find $\text{Var}(Y)$.

Solution. $X_1, X_2, \cdots, X_n$ be the fair coin flips.

(a) $P(X_1 \oplus X_2 = 0) = P(X_1 = X_2)$

$= P_{X_1X_2}(0,0) + P_{X_1X_2}(1,1) = 1/2$

$P(X_1 \oplus X_2 = 1) = 1/2$.

So $P(Y_i = 0) = P(Y_i = 1) = 1/2$ since $X_i = X_j + X_k$ for some $j \neq k$.

(b) Consider $Y_i = X_1 \oplus X_2$, $Y_j = X_2 \oplus X_3$, $Y_k = X_1 \oplus X_3$.

Y_i, Y_j, and Y_k are not independent.

$P(Y_i = 1, Y_j = 1, Y_k = 1)$

$= 0 \neq 1/8 = P(Y_i = 1) P(Y_j = 1) P(Y_k = 1)$.

(c) $E[Y_i] = 1/2$.

$E[Y_i Y_j] = E[(X_i \oplus X_j)(X_k \oplus X_l)]$

for $i \neq j$, $k \neq l$, $\{i,j\} \neq \{k,l\}$

If $\{i,j\} \cap \{k,l\} = \phi$, $E[Y_i Y_j] = E[X_i \oplus X_j] E[X_k \oplus X_l]$

$= 1/4$.

Suppose $|\{i,j\} \cap \{k,l\}| = 1$.

$$E[(x_1 \oplus x_2)(x_2 \oplus x_3)]$$

$$= \sum_{x_1 x_2 x_3} (x_1 \oplus x_2)(x_2 \oplus x_3)\, p_{x_1}(x_1)\, p_{x_2}(x_2)\, p_{x_3}(x_3)$$

$$= \frac{1}{8} \sum_{x_1 x_2 x_3} (x_1 \oplus x_2)(x_2 \oplus x_3)$$

0	1	0
1	0	1

$$= \frac{2}{8} = \frac{1}{4}$$

$$\therefore E[x_i x_j] = E[x_i]\, E[x_j] = \frac{1}{4}.$$

(d) $\mathrm{Var}(Y) = E[Y^2] - (E[Y])^2$

$$= \sum_{i=1}^{m} E[x_i^2] + \sum_{i \neq j} E[x_i x_j] - \left(m \tfrac{1}{2}\right)^2$$

$$= \frac{m}{2} + \frac{m(m-1)}{4} - \frac{m^2}{4} = \frac{n(n-1)}{8}.$$

Q) We say that α is a median of a RV x if $P(x \leq \alpha) \geq \frac{1}{2}$ and $P(x \geq \alpha) \geq \frac{1}{2}$.

It is possible for the median to be non-unique, with all values in an interval satisfying the definition.

(a) Let $x \in \{0, 1, 2\}$ with probabilities p_0, p_1, and p_2, respectively. Find the median of x for each of the cases below.

(i) $p_0 = 0,2$ $p_1 = 0,4$ $p_2 = 0,4$

(ii) $p_0 = 0,2$ $p_1 = 0,2$ $p_2 = 0,6$

(iii) $p_0 = 0,2$ $p_1 = 0,3$ $p_2 = 0,5$.

(b) Suppose x is a continuous RV with PDF f_x s.t. $f_x(x) = \begin{cases} 0,5 & \text{for } 0 \leq x \leq 0,5 \\ 0 & \text{for } 0,5 < x \leq 1 \end{cases}$.

We know that $f_x(x) > 0$ for all $x > 1$, and $f_x(x) = 0$ for all $x < 0$, but is otherwise unknown. Find the median.

Solution: (a) (i) $\alpha = 1$ is the median.

$$P(X \le 1) = 0.6 \ge \frac{1}{2}$$

$$P(X \ge 1) = 0.8 \ge \frac{1}{2}$$

(ii) $\alpha = 2$ is the median.

$$P(X \le 2) = 1 \ge \frac{1}{2}$$

$$P(X \ge 2) = 0.6 \ge \frac{1}{2}$$

(iii) Let $\alpha \in [1, 2]$.

$$P(X \le \alpha) = \begin{cases} 0.5 & \text{if } \alpha \in [1, 2) \\ 1 & \text{if } \alpha = 2 \end{cases}$$

$$\ge \frac{1}{2}$$

$$P(X \ge \alpha) = \begin{cases} 0.5 & \text{if } \alpha \in (1, 2] \\ 0.8 & \text{if } \alpha = 1 \end{cases} \quad \ge \frac{1}{2}$$

(b) $P(X \le 0.5) = 0.5$

Let $\alpha \in [0.5, 1]$.

$$P(X \le \alpha) = 0.5 \ge \frac{1}{2}$$

$$P(X \ge \alpha) = \int_{1}^{\infty} f_X(x)\,dx = \frac{1}{2} \ge \frac{1}{2}$$

Q) Suppose x and y are two independent RVs such that

$E[x^4] = 2$ $E[y^2] = 1$ $E[x^2] = 1$ and $E[y] = 0$,

Find $\text{Var}(x^2y)$.

$$\text{Var}(x^2y) = E[x^4y^2] - (E[x^2y])^2$$

$$= 2.1 - (1.0)^2 = 2.$$

Lecture 14

(25 September 2023)

Geometric and Exponential CDFs

CDF defined for any type of RV provides a convenient means for exploring the relations between continuous and discrete random variables. We explore the relation between geometric and exponential RVs.

Let X be a geometric RV with success probability p i.e., X is the no. of trials until the first success in a sequence of independent Bernoulli trials where the probability of success at each trial is p.

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

$$F_X^G(n) = \sum_{k=1}^{n} P_X(k) = p \cdot \frac{1-(1-p)^n}{p} = 1-(1-p)^n.$$

If X is exponential RV

$$F_X^E(x) = \begin{cases} \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x} & \text{for } x > 0. \\ 0 & \text{for } x \le 0. \end{cases}$$

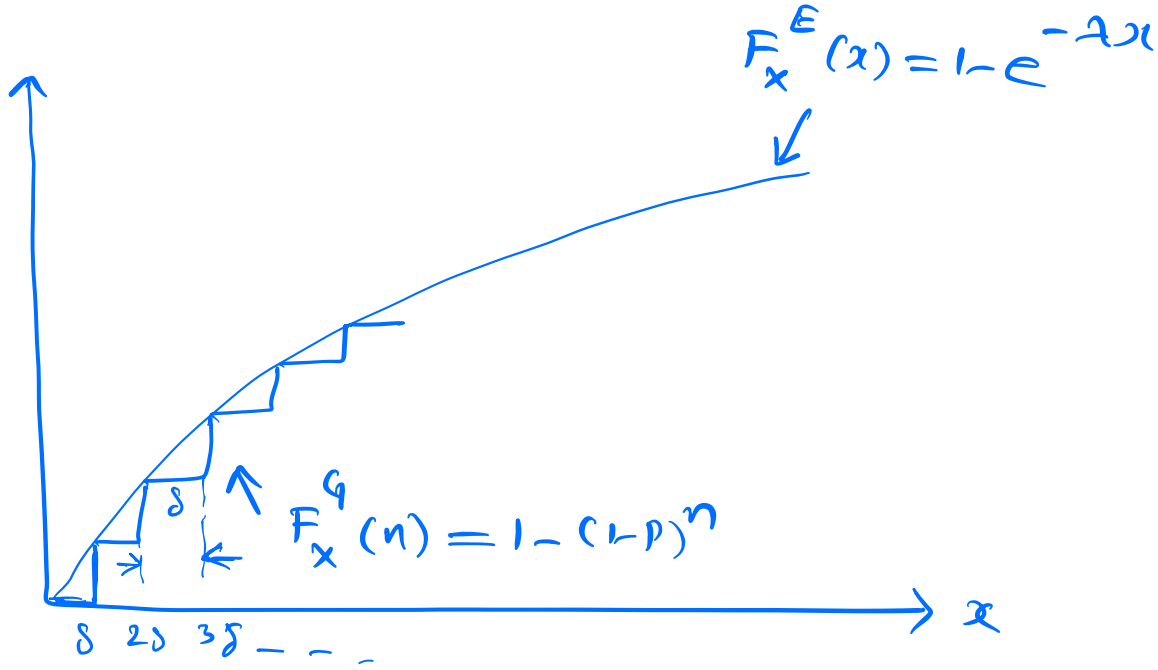
Consider the values of $F_x^E(x)$ at $x = n\delta$ $n = 1, 2, \ldots$

$$1 - e^{-\lambda n\delta} = 1 - (1-p)^n$$

$$e^{-\lambda\delta} = 1 - p \Rightarrow \delta = -\ln(1-p)/\lambda .$$

Then we see that the values of the exponential and the geometric CDFs are equal whenever $x = n\delta$ with $n = 1, 2, \ldots$ i.e.,

$$F_x^E(n\delta) = F_x^G(n) \quad n = 1, 2, \ldots$$



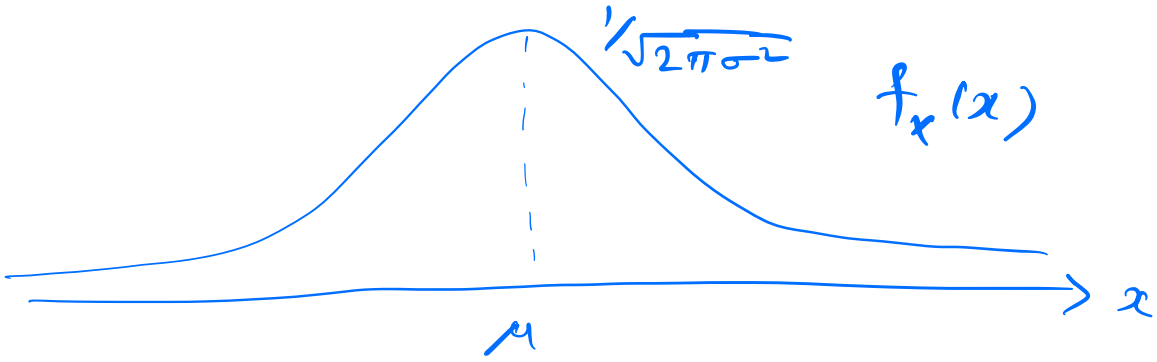
suppose now we toss a biased coin very quickly (every δ seconds where $\delta \ll 1$) with a small probability of heads (equal to $p = 1 - e^{-\lambda\delta}$). Then the first time to obtain a head (a geometric random variable with parameter p) is a close approximation to an exponential r.v. with parameter λ in the sense that the corresponding CDFs are very close to each other as shown in the above figure.

Gaussian Random Variables

A continuous RV X is said to be Gaussian or normal if it has a PDF of the form

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

where $\mu \in \mathbb{R}$, $\sigma \in (0, \infty)$.



Is it a valid PDF ?

$$\int_{-\infty}^{\infty} f_x(x)\,dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\,dx$$

Let $t = \dfrac{x-\mu}{\sigma}$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt = I$$

$$I^2 = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{-y^2/2}\,dx\,dy$$

$$= \frac{1}{2\pi} \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy$$

$x = r\cos\theta$ $y = r\sin\theta$ gives

$$\frac{1}{2\pi}\int_{r=0}^{\infty}\int_{\theta=0}^{2\pi} e^{-r^2/2}\, r\, dr\, d\theta$$

$$= \frac{1}{2\pi}(2\pi)\int_{r=0}^{\infty} e^{-r^2/2}\, r\, dr$$

$$= -e^{-r^2/2}\Big|_{r=0}^{\infty} = 1.$$

$$\therefore I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1.$$

$$E[x] = \int_{-\infty}^{\infty} x\, \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}\, dx$$

$$= \int_{-\infty}^{\infty} (\sigma t + \mu)\, \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}\, dt$$

$$= \mu + \underbrace{\int_{-\infty}^{\infty} \sigma\, t\, \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}\, dt}$$

$$= \frac{\sigma}{\sqrt{2\pi}}\, -e^{-t^2/2}\Big|_{-\infty}^{\infty} = 0$$

So mean $= E[x] = \mu$.

Intuitively note that the PDF is symmetric around μ, so the mean can only be μ.

$$\text{Var}(x) = E[x^2] - (E[x])^2$$

$$E[x^2] = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \, dx$$

$$= \int_{-\infty}^{\infty} (\sigma t + \mu)^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt$$

$$= \mu^2 + \sigma^2 \int_{-\infty}^{\infty} t^2 \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt + 2\mu\sigma \cdot 0$$

$$= \mu^2 + \sigma^2 \left(-t \cdot e^{-t^2/2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} 1 \cdot \frac{e^{-t^2/2}}{\sqrt{2\pi}} \, dt \right)$$

$$= \mu^2 + \sigma^2$$

$\therefore \text{Var}(x) = \sigma^2$.

A normal RV has several special properties.

Theorem. If x is a normal random variable with mean μ and variance σ^2, and if $a \neq 0$, b are scalars then the random variable

$y = ax + b$ is also normal with mean $E[y] = aE[x] + b$ $\quad var(y) = a^2\sigma^2$.

Before proving this theorem let us find the PDF of a linear function of any RV.

Let x be a continuous RV with PDF f_x, and let $y = ax + b$.

$$P(Y \leq y) = P(ax + b \leq y)$$

$$= P(ax \leq y - b)$$

$$= \begin{cases} P\left(x \leq \frac{y-b}{a}\right) & \text{if } a > 0 \\ P\left(x \geq \frac{y-b}{a}\right) & \text{if } a < 0 \end{cases}$$

$$f_Y(y) = \frac{dF_Y(y)}{dy}$$

$$= \begin{cases} \frac{1}{a} f_X\left(\frac{y-b}{a}\right) & \text{if } a > 0 \\ \frac{-1}{a} f_X\left(\frac{y-b}{a}\right) & \text{if } a < 0 \end{cases}$$

$$X \sim f_X \implies aX+b \sim \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Proof of Theorem.

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

$$= \frac{1}{|a|} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{y-b}{a}-\mu\right)^2/2\sigma^2}$$

$$= \frac{1}{\sqrt{2\pi|a|^2\sigma^2}} e^{-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}}$$

This is a normal PDF with mean $a\mu+b$ and variance $a^2\sigma^2$. Thus $Y=aX+b$ is a normal RV.

A Linear Function of an Exponential RV.

Suppose x is an exponential RV with PDF

$$f_x(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{o.w.} \end{cases}$$

Let $y = ax + b$.

$$f_y(y) = \frac{1}{|a|} f_x\left(\frac{y-b}{a}\right)$$

$$= \begin{cases} \frac{\lambda}{|a|} e^{-\lambda \frac{y-b}{a}}, & \text{if } \frac{y-b}{a} \geq 0 \\ 0, & \text{o.w.} \end{cases}$$

If $b=0$ and $a>0$ then y is an exponential RV with parameter λ/a. In general y need not be exponential.

— A normal RV y with zero mean and unit variance is said to be a standard normal RV. Its CDF is denoted by Φ:

$$\Phi(y) = P(Y \leq y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-t^2/2} dt.$$

$$\Phi(-y) = P(Y \le -y)$$

$$= P(Y \ge y) = 1 - P(Y \le y)$$

$$= 1 - \Phi(y).$$

- Let X be a normal RV with mean μ and variance σ^2. We "standardize" X by defining a new RV Y given by

$$Y = \frac{X - \mu}{\sigma}.$$

$$E[Y] = 0 \quad \text{Var}(Y) = 1.$$

Example. $X \sim N(\mu = 60, \sigma^2 = 20^2)$.

$$P(X \ge 80) = P\left(\frac{X - 60}{20} \ge \frac{80 - 60}{20} = 1\right)$$

$$= P(Y \ge 1) = \Phi(1).$$

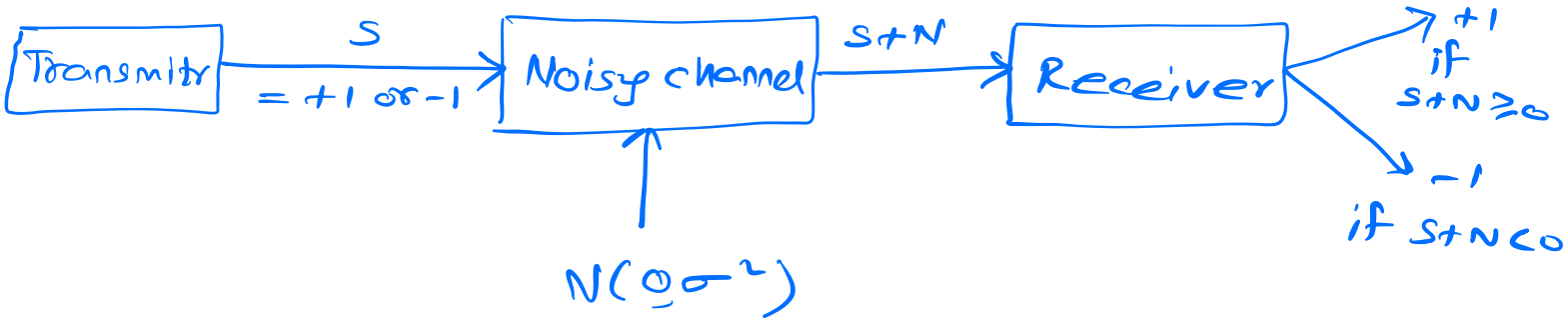
In general

$$P(X \le x) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

Normal RVs are often used in signal processing and communications to model noise.

Example (signal Detection).

A binary message is transmitted as a signal S, which is either $+1$ or -1.



what is the probability of error?

An error occurs whenever -1 is transmitted and the Noise N is at least 1 so that $S+N = -1+N \geq 0$ or whenever $+1$ is transmitted and the noise N is smaller then -1 so that $S+N = 1+N < 0$.

when $S=-1$ $P(N \geq 1)$ is the error probability.

$$P(N \geq 1) = 1 - P(N < 1) = 1 - P\left(\frac{N}{\sigma} < \frac{1}{\sigma}\right)$$

$$= 1 - \Phi\left(\frac{1}{\sigma}\right).$$

when $S=+1$ $P(N \leq -1) = P(N \geq 1)$

$$= 1 - \Phi\left(\frac{1}{\sigma}\right).$$

Joint PDFs of Multiple Random Variables

We say that two random variables associated with the same experiment are jointly continuous if there exists a non-negative function f_{XY}, called as joint PDF, such that

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

for every subset B of $\mathbb{R}^2$.

Exercise. If X and Y are jointly continuous, prove that they are individually continuous also.

Lecture 15

(5 September 2023)

Joint PDFs of Multiple Random Variables

We say that two random variables associated with the same experiment are jointly continuous if there exists a non-negative function f_{XY}, called as joint PDF, such that

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

for every subset B of $\mathbb{R}^2$.

In particular, when B is a rectangle

$$P(a \le X \le b,\ c \le Y \le d) = \int_{y=c}^{d} \int_{x=a}^{b} f_{XY}(x,y)\, dx\, dy.$$

$$B = \mathbb{R}^2 \Rightarrow \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_X(x,y)\, dx\, dy = 1.$$

To interpret the joint PDF, we let δ be a small positive number and consider the probability of a small rectangle.

$$P(a \le X \le a+\delta \quad c \le Y \le c+\delta)$$

$$= \int_{c}^{c+\delta} \int_{a}^{a+\delta} f_{XY}(x,y)\,dx\,dy \approx f_{XY}(a,c)\,\delta^2$$

so we can view $f_{XY}(a,c)$ as the "probability per unit area" in the vicinity of (a,c).

$$P(X \in A) = P(X \in A \text{ and } Y \in (-\infty,\infty))$$

$$= \int_{x \in A} \int_{y=-\infty}^{\infty} f_{XY}(x,y)\,dy\,dx.$$

If x and y are jointly continuous, they are individually continuous.

Comparing with $P(X \in A) = \int_A f_X(x)\,dx$

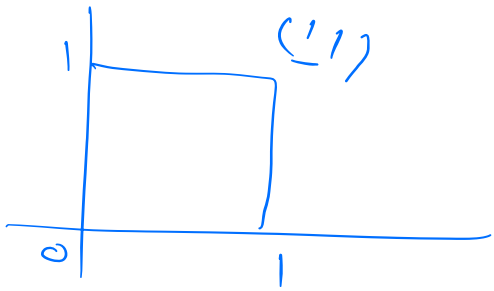
marginal pdf f_X of x is given by

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

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

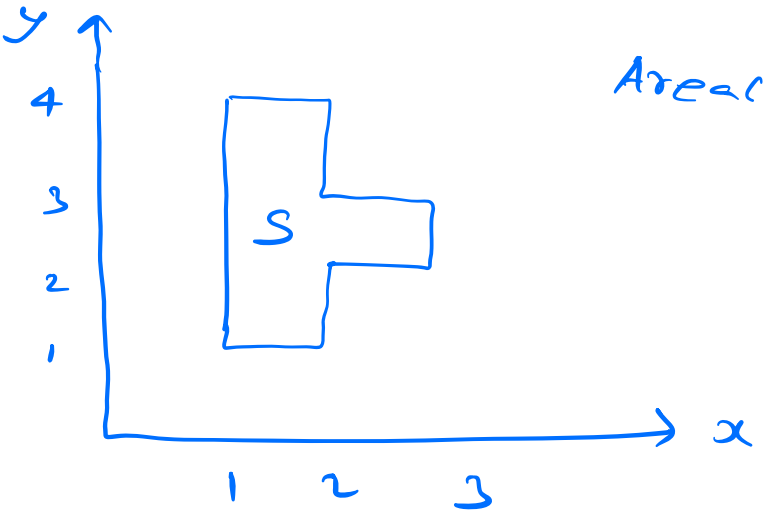
Example (Two-Dimensional Uniform PDF).

$$f_{XY}(x,y) = \begin{cases} c & \text{if } 0 \le x \le 1 \text{ and } 0 \le y \le 1 \\ 0 & \text{o.w.} \end{cases}$$



where c is a constant.

Example.



Area(S) = 4

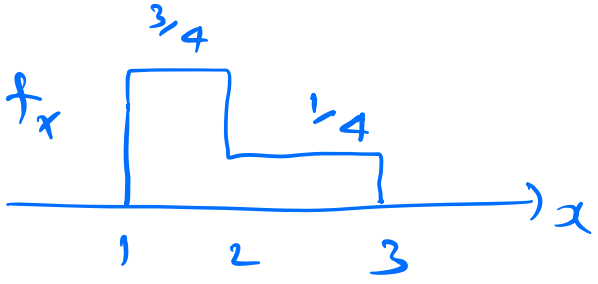
$$f_{XY}(x,y) = \begin{cases} 1/4 & \text{if } (x,y) \in S \\ 0 & \text{o.w.} \end{cases}$$

Find f_X, f_Y.

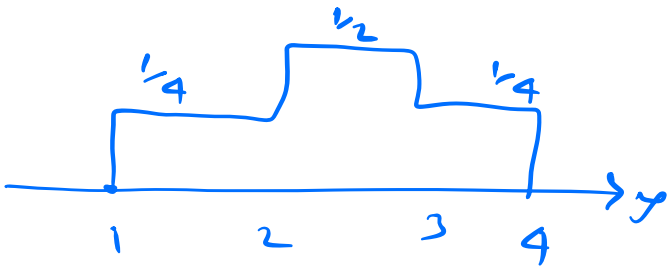
$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy$$

$$= \begin{cases} \int_1^4 \frac{1}{4}\, dy & x \in [1:2] \\ \int_2^3 \frac{1}{4}\, dy & x \in [2:3] \end{cases}$$

$=$



$f_y =$



Joint CDFs. If x and y are two RVs associated with the same experiment we define their joint CDF by

$$F_{xy}(x,y) = P(X \le x,\ Y \le y).$$

For continuous RVs x and y

$$F_{xy}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{xy}(s,t)\, dt\, ds.$$

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

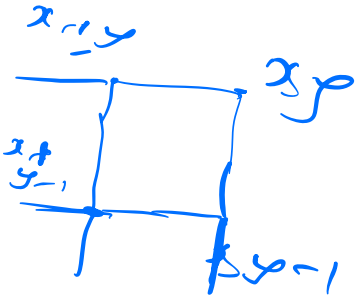
$$P(x_1 < x \le x_2,\ y_1 < y \le y_2)$$

$$= F_{xy}(x_2, y_2) - F_{xy}(x_1, y_2) - F_{xy}(x_2, y_1) + F_{xy}(x_1, y_1)$$

For discrete RVs x and y

$$F_{xy}(x, y) = \sum_{l \le x} \sum_{k \le y} P_{xy}(l, k).$$

$$P_{xy}(x, y) = F_{xy}(x, y) - F_{xy}(x-1, y) - F_{xy}(x, y-1) + F_{xy}(x-1, y-1)$$



Example. Let x and y be described by a uniform PDF on the unit square.

$$F_{xy}(x, y) = P(x \le x, y \le y) = xy \quad \text{for } 0 \le x, y \le 1.$$

$$F_{xy}(x, y) = x \quad \text{if } 0 \le x \le 1 \text{ and } y \ge 1.$$

$$f_{xy}(x, y) = \frac{\partial^2 F_{xy}(x, y)}{\partial x \partial y} = 1 \quad \text{for } 0 \le x, y \le 1.$$

Properties of Joint CDF

(1) $\lim_{x \to \infty} F_{XY}(x,y) = F_Y(y)$,

$\lim_{y \to \infty} F_{XY}(x,y) = F_X(x)$

$\lim_{\substack{x \to \infty \\ y \to \infty}} F_{XY}(x,y) = 1$.

(2) $\lim_{x \to -\infty} F_{XY}(x,y) = 0 = \lim_{y \to -\infty} F_{XY}(x,y)$.

(3) If $x_1 \le x_2, y_1 \le y_2$ then

$$F_{XY}(x_1,y_1) \le F_{XY}(x_2,y_2).$$

(4) $\lim_{\substack{\varepsilon \to 0^+ \\ \delta \to 0^+}} F_{XY}(x+\varepsilon, y+\delta) = F_{XY}(x,y)$.

Expected value Rule

If X and Y are jointly continuous RVs and g is some function then $Z = g(X,Y)$ is

also a RV,

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)\,dx\,dy.$$

$$E[aX+bY+c] = aE[X] + bE[Y] + c.$$

More than Two Random Variables

X, Y and Z are jointly continuous if

$\exists$ non-negative f_{XYZ} s.t.

$$P((X,Y,Z) \in B) = \int_{(x,y,z)\in B} f_{XYZ}(x,y,z)\,dx\,dy\,dz$$

for any set $B \subseteq \mathbb{R}^3$.

$\Rightarrow$ X and Y are jointly continuous with joint PDF

$$f_{XY}(x,y) = \int_{-\infty}^{\infty} f_{XYZ}(x,y,z)\,dz.$$

$$f_X(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XYZ}(x,y,z)\,dy\,dz.$$

$$E\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i E[X_i].$$

Conditioning a RV on an Event

The conditional PDF of a continuous RV X given an event A with $P(A) > 0$ is defined as a non-negative function $f_{X|A}$ that satisfies

$$P(X \in B \mid A) = \int_B f_{X|A}(x)\, dx,$$

for any $B \subseteq \mathbb{R}$ and $A \subseteq \Omega$.

$$B = \mathbb{R} \Rightarrow \int_{-\infty}^{\infty} f_{X|A}(x)\, dx = 1.$$

Suppose A is of the form $A = \{X \in C\}$.

$$P(X \in B \mid X \in C) = \frac{P(X \in B \cap C)}{P(X \in C)}$$

$$= \frac{1}{P(X \in C)} \int_{B \cap C} f_X(x)\, dx$$

$$= \int_B \frac{f_X(x)\, 1\{x \in C\}}{P(X \in C)}\, dx$$

so $f_{X|X\in C}(x) = \begin{cases} \frac{f_X(x)}{P(X\in C)}, & \text{if } x\in C \\ 0, & \text{o.w.} \end{cases}$

Conditional CDF

$$F_{X|A}(x) = P(X \le x \mid A).$$

Theorem. Let $A_1, A_2, \ldots, A_n$ be disjoint sets that form a partition of the sample space, and assume that $P(A_i) > 0$ for all i. Then

$$f_X(x) = \sum_{i=1}^{n} P(A_i) f_{X|A_i}(x).$$

Proof. $F_X(x) = P(X \le x)$

$$= \sum_{i=1}^{n} P(A_i) F_{X|A_i}(x)$$

Differentiating w.r.t. x, we get

$$f_X(x) = \sum_{i=1}^{n} P(A_i) f_{X|A_i}(x).$$

Lecture 16

(9 September 2023)

Conditioning one RV on another

Consider a f_{XY}. For any y with $f_Y(y) > 0$ the conditional PDF of x given $Y = y$ is defined by

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

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

To interpret the conditional probability consider

$$P(x \le X \le x+\delta_1 \mid y \le Y \le y+\delta_2)$$

$$= P(x \le X \le x+\delta_1,\ y \le Y \le y+\delta_2) \Big/ P(y \le Y \le y+\delta_2)$$

$$\approx \frac{f_{XY}(x,y)\,\delta_1\delta_2}{f_Y(y)\,\delta_2}.$$

In view of the above

Define $P(X \in A \mid Y = y) = \int_A f_{X|Y}(x|y)\,dx$.

Interpretation:

$$P(X \in A \mid Y = y) := \lim_{\delta \to 0} P(X \in A \mid y \le Y \le y+\delta)$$

$$= \lim_{\delta \to 0} \frac{P(X \in A,\ y \le Y \le y+\delta)/\delta}{P(y \le Y \le y+\delta)/\delta}$$

$$= \lim_{\delta \to 0} \frac{\int_{x \in A} \int_{y}^{y+\delta} f_{XY}(x,t)\,dt\,dx}{\delta} \Big/ \int_{y}^{y+\delta} f_Y(y)\,dy$$

$$= \int_{x \in A} f_X(x) \left[\lim_{\delta \to 0} \int_{y}^{y+\delta} f_{Y|X}(t|x)\,dt / \delta \right] dx \Big/ \lim_{\delta \to 0} \int_{y}^{y+\delta} f_Y(y)\,dy$$

$$= \int_{x \in A} f_X(x) \frac{d}{dy} F_{Y|X}(y|x) \Big/ \frac{d}{dy} f_Y(y)\,dx$$

$$= \int_{x \in A} \frac{f_X(x) f_{Y|X}(y|x)}{f_Y(y)}\,dx.$$

Conditional Expectation

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

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

Expected value Rule:

$$E[g(X)|A] = \int_{-\infty}^{\infty} g(x)\, f_{X|A}(x)\, dx.$$

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

Total Expectation Theorems.

i) Let $A_1, A_2, \ldots, A_n$ be disjoint events that form a partition of the sample space and let $P(A_i) > 0$ for all i. Then

$$E[X] = \sum_{i=1}^{n} P(A_i) E[X|A_i].$$

proof. $$f_X(x) = \sum_{i=1}^{n} P(A_i) f_{X|A_i}(x)$$

multiply both sides by x and then integrate to

Get $E[X] = \sum_{i=1}^{n} P(A_i) E[X|A_i]$.

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

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

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

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

— The total expectation theorem can often be used to calculate mean and variance.

Example. Mean and variance of a piecewise constant PDF.

$$f_X(x) = \begin{cases} 1/3 & \text{if } 0 \le x \le 1 \\ 2/3 & \text{if } 1 \le x \le 2 \\ 0 & \text{o.w.} \end{cases}$$

$A_1 = \{x \text{ lies in } [0,1]\}$ $A_2 = \{x \text{ lies in } (1,2]\}$.

$P(A_1) = \frac{1}{3}$, $P(A_2) = \frac{2}{3}$.

$E[X] = P(A_1)E[X|A_1] + P(A_2)E[X|A_2]$

$f_{X|A_1}$, $f_{X|A_2}$ are uniform.

Recall a $U \sim \text{Uniform}[a,b] \Rightarrow E[U] = \frac{a+b}{2}$,

$E[U^2] = \frac{a^2+ab+b^2}{3}$.

$E[X|A_1] = \frac{1}{2}$, $E[X|A_2] = \frac{3}{2}$,

$E[X^2|A_1] = \frac{1}{3}$ $E[X^2|A_2] = \frac{7}{3}$.

$E[X] = \frac{7}{6}$ $E[X^2] = \frac{15}{9}$.

$\text{Var}(X) = E[X^2] - E[X]^2 = \frac{11}{36}$.

Independence

Two continuous RVs are independent if

$f_{XY}(x,y) = f_X(x) f_Y(y)$ $\forall x, y$.

This is same as

$f_{X|Y}(x|y) = f_X(x)$ for all y with $f_Y(y) > 0$ and for all x.

Example. Independent Normal RVs.

Let x and y be independent Gaussian RVs with means μ_x, μ_y and variances σ_x^2, σ_y^2 respectively.

$$f_{xy}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}}.$$

Theorem. Two random variables x and y (continuous or discrete) are independent if and only if

$$F_{xy}(x,y) = F_x(x) F_y(y), \text{ for all } x, y.$$

proof. We prove for the continuous case.

Let x and y are independent.

$$F_{xy}(x,y) = P(x \le x, y \le y)$$

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

$$= \int_{-\infty}^{x} f_x(x) dx \int_{-\infty}^{y} f_y(y) dy = F_x(x) F_y(y).$$

For the other direction let

$$F_{xy}(x,y) = F_x(x) F_y(y).$$

On taking second order mixed partial derivative

$$\frac{\partial^2 F_{xy}(x,y)}{\partial x \partial y} = \frac{\partial F_x(x)}{\partial x} \cdot \frac{\partial F_y(y)}{\partial y}$$

$$\Rightarrow f_{xy}(x,y) = f_x(x) f_y(y).$$

Exercise. Prove the above theorem for the discrete case.

- If x and y are independent, then

$$E[xy] = E[x] E[y].$$

Theorem. If x and y are jointly continuous independent RVs. For any two functions g & h

$$E[g(x)h(y)] = E[g(x)] E[h(y)].$$

Proof. It suffices to show that g(x) &

$h(y)$ are independent.

Let $x' = g(x)$, $y' = h(y)$.

$$F_{x'y'}(x'y') = P(g(x) \le x', h(y) \le y')$$

$$= \int_{\{x : g(x) \le x'\}} \int_{\{y : h(y) \le y'\}} f_{xy}(x,y)\, dx\, dy$$

$$= \int_{x : g(x) \le x'} f_x(x)\, dx \cdot \int_{y : h(y) \le y'} f_y(y)\, dy$$

$$= P(g(x) \le x')\ P(h(y) \le y')$$

$$= F_{x'}(x')\, F_{y'}(y').$$

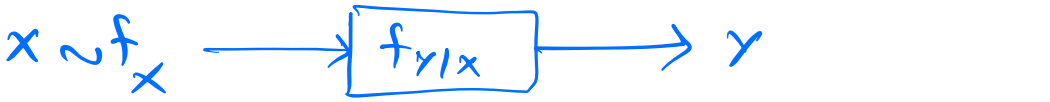
— If x and y are independent, then

$$Var(x+y) = Var(x) + Var(y).$$

Lecture 17

(12 September 2023)

The Continuous Bayes' Rule



Unknown RV with known PDF

Observed RV

Goal: Infer about X?

The information provided by the event $\{Y=y\}$ is captured by the conditional PDF $f_{X|Y}(x|y)$.

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

$$\Rightarrow f_{X|Y}(x|y) = \frac{f_X(x) f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(t) f_{Y|X}(y|t)\, dt}$$

Inference about a Discrete RV:

$$P(A|Y=y) = \lim_{\delta \to 0} P(A | Y \in [y, y+\delta])$$

$$= \lim_{\delta \to 0} \frac{P(A)\, P(Y \in [y, y+\delta] \mid A)/\delta}{P(Y \in [y, y+\delta])/\delta}$$

$$= \frac{P(A)\, f_{Y|A}(y)}{f_Y(y)}$$

Also $f_Y(y) = P(A) f_{Y|A}(y) + P(A^c) f_{Y|A^c}(y)$.

So

$$P(A \mid Y=y) = \frac{P(A) f_{Y|A}(y)}{P(A) f_{Y|A}(y) + P(A^c) f_{Y|A^c}(y)}$$

Let X be a discrete RV and Y be a continuous RV. The above equality gives

$$X \sim P_X \longrightarrow \boxed{f_{Y|X}(y|x)} \longrightarrow Y$$

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

$$= \frac{P_x(x)\, f_{y|x}(y|x)}{\sum_{x=0}^{M-1} P_x(x)\, f_{y|x}(y|x)}$$, where $\sum_{x=0}^{M-1} P_x(x) = 1$.

Goal: To estimate the hypothesis x that lead to an observation y.

- A test $\hat{x}(y)$ is a decision rule or a deterministic function of the observation y.
- $P_{x|y}(x|y)$ is the probabability that hypothesis x is correct, i.e., the probability that $X = x$, conditional on observation y.

$P_{x|y}(x|y)$ is called 'a posteriori probability'.

Consider the decision rule that maximizes this a posteriori probability.

$$\hat{x}_{MAP}(y) = \arg\max_x P_{x|y}(x|y) \quad \text{(MAP rule)},$$

where $\arg\max_x$ means the argument $x \in \{0, 1, \ldots, M-1\}$ that maximizes the function.

Using the Baxes' law,

$$\hat{x}_{MAP}(y) = \arg\max_x \frac{p_x(x) f_{y|x}(y|x)}{f_y(y)}$$

$$= \arg\max_x p_x(x) f_{y|x}(y|x).$$

When multiple hypotheses achieve the maximum, we arbitrarily choose the largest maximizing x.

- For any test A, $p_{x|y}(\hat{x}_A(y)|y)$ is the probability that $\hat{x}_A(y)$ is the correct decision when test A is used on observation y. Since $\hat{x}_{MAP}(y)$ maximizes the probability of correct decision, we have

$$p_{x|y}(\hat{x}_{MAP}(y)|y) \geq p_{x|y}(\hat{x}_A(y)|y)$$

for all A and y.

- Probability of correctness for a test A:

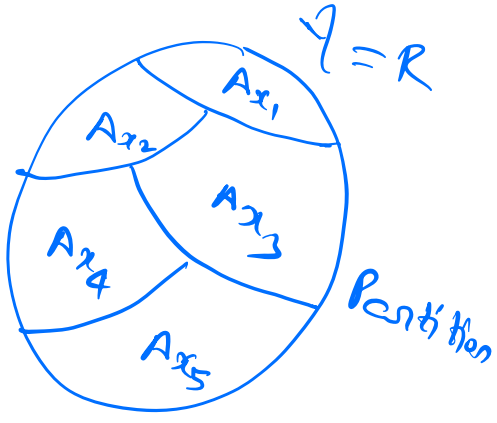
$$P(\hat{x}_A(y) = x).$$

Theorem. The MAP rule maximizes the probability of correct decision conditional on each observed sample y. It also maximizes the overall probability of correct decision defined above.

proof. $A_x = \{y : \hat{x}_A(y) = x\}$ be the set of observations y that test A maps to hypothesis x.

$$P(\hat{x}_A(y) = x) = \sum_{x=0}^{M-1} P_x(x)\, P(\hat{x}_A(y) = x \mid X = x)$$

$$= \sum_{x=0}^{M-1} P_x(x)\, P(Y \in A_x \mid X = x)$$



$$= \sum_{x=0}^{M-1} P_x(x) \int_{y \in A_x} f_{Y|X}(y|x)\, dy$$

$$= \int_{y=-\infty}^{\infty} P_x(\hat{x}_A(y))\, f_{Y|X}(y|\hat{x}_A(y))\, dy$$

$$= \int_{y=-\infty}^{\infty} P_{X|Y}(\hat{x}_A(y)|y)\, f_Y(y)\, dy \longrightarrow (1)$$

Similarly

$$P(\hat{x}_{MAP}(Y)=x)=\int_{y=-\infty}^{\infty} P_{X|Y}(\hat{x}_{MAP}(y)|y)\, f_Y(y)\, dy \quad \rightarrow (2)$$

Now since

$$P_{X|Y}(\hat{x}_{MAP}(y)|y) \geq P_{X|Y}(\hat{x}_A(y)|y) \text{ for all } A, y$$

averaging gives

$$(2) \geq (1), \text{ i.e. } P(\hat{x}_{MAP}(Y)=x) \geq P(\hat{x}_A(Y)=x).$$

(Revisiting) Binary MAP detection.

Let $P_X(0)=P_0$, $P_X(1)=P_1$, $0<P_0, P_1$, $P_1+P_0=1$.

Let Y be a continuous RV with conditional PDF $f_{Y|X}(y|x)$.

$$f_Y(y)=P_0 f_{Y|X}(y|0)+P_1 f_{Y|X}(y|1)>0.$$

MAP rule for a fixed y

$$P_{X|Y}(1|y) \underset{\hat{x}(y)=0}{\overset{\hat{x}(y)=1}{\gtrless}} P_{X|Y}(0|y)$$

Equivalently using Bayes' rule

$$\frac{f_{Y|X}(y|1)P_X(1)}{f_Y(y)} \underset{\hat{x}(y)=0}{\overset{\hat{x}(y)=1}{\gtrless}} \frac{f_{Y|X}(y|0)P_X(0)}{f_Y(y)},$$

equivalently

$$\Lambda(y) = \frac{f_{Y|X}(y|1)}{f_{Y|X}(y|0)} \underset{\hat{x}(y)=0}{\overset{\hat{x}(y)=1}{\gtrless}} \frac{P_0}{P_1} = \eta.$$

↑ Likelihood ratio

[Threshold rule]

- Note that if the a priori probability P_0 is increased then the threshold increases and the set of y for which hypotheses 0 is chosen increases; this corresponds to our intuition - the greater our initial conviction that x is 0 the stronger the evidence required to change our mind.

- A special case of the above threshold rule is when $P_0 = P_1$. It is called maximum Likelihood (ML) test. The ML test is often used when P_0 and P_1 are unknown.

The partition of the space of observed sample values is given by

$$A_1 = \{y : \Lambda(y) \geq \eta\} \qquad A_0 = \{y : \Lambda(y) < \eta\}$$

$$\downarrow \qquad\qquad \downarrow$$

$$\hat{x} = 1 \qquad\qquad \hat{x} = 0.$$

We can compute the probability of error $P(\text{Error} \mid X = x)$.

$$P(\text{Er} \mid X = 0) = P(\hat{x}(Y) = 1 \mid X = 0)$$

$$= P(Y \in A_1 \mid X = 0)$$

$$= \int_{y \in A_1} f_{Y|X}(y|0)\, dy.$$

Similarly $P(\text{Er} \mid X = 1) = \int_{y \in A_0} f_{Y|X}(y|1)\, dy.$

The overall probability of error

$$= P_0 P(\text{Er} \mid X = 0) + P_1 P(\text{Er} \mid X = 1).$$

Example. (Abstraction of a digital communication system)

$P_X(b) = P_1$ $P_X(-b) = P_0$. $Y = X + Z$, $Z \sim N(0, \sigma^2)$,

X and Z are independent.

$$P(Y \in B \mid X = x) = P(x + Z \in B \mid X = x)$$

$$= P(Z + x \in B)$$

$\therefore Y|X=b \sim N(b, \sigma^2)$ $\quad Y|X=-b \sim N(-b, \sigma^2)$,

$$f_{Y|X}(y|b) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-b)^2/2\sigma^2}$$

$$f_{Y|X}(y|-b) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y+b)^2/2\sigma^2}.$$

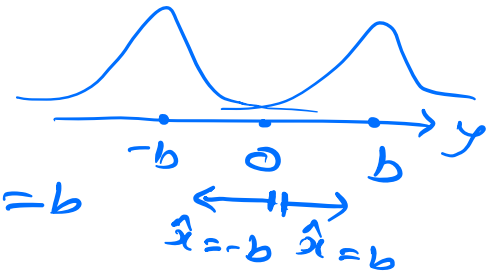
MAP rule:

$$\Lambda(y) = \frac{f_{Y|X}(y|b)}{f_{Y|X}(y|-b)} = e^{\frac{(y+b)^2 - (y-b)^2}{2\sigma^2}}$$

$$= e^{2yb/\sigma^2} \quad \begin{matrix} \hat{x}(y)=b \\ \geq \\ < \\ \hat{x}(y)=-b \end{matrix} \quad \frac{p_0}{p_1} = \eta.$$

$$\Rightarrow y \quad \begin{matrix} \hat{x}(y)=b \\ \geq \\ < \\ \hat{x}(y)=-b \end{matrix} \quad \frac{\sigma^2}{2b} \log \eta.$$

If $b=1$ and $p_0=p_1$ this recovers the example on signal detection we have seen in an earlier lecture. Then



ML rule maps $y \geq 0 \rightarrow \hat{x}=b$

$y<0 \rightarrow \hat{x}=-b.$

Coming back to the case with $b>0$

$$P(Err|X=b) = P(Y<0 \mid X=b)$$

$$= P\left(\frac{Y-b}{\sigma} < \frac{-b}{\sigma} \mid X=b\right)$$

$$= P\left(N < \frac{-b}{\sigma}\right)$$

$$= P\left(N \geq \frac{b}{\sigma}\right).$$

$$P(Err|X=-b) = P\left(N > \frac{b}{\sigma}\right)$$

$$= 1 - P\left(N \leq \frac{b}{\sigma}\right).$$

$$P(Error) = P\left(N \geq \frac{b}{\sigma}\right).$$

Lecture 18

(16 October 2023)

Functions of Random Variables

$Y = g(X)$, i.e., $Y(\omega) = g(X(\omega))$.

For the discrete case

$$P_Y(y) = \sum_{x : g(x) = y} P_X(x).$$

For the continuous case we follow the two-step procedure outlined below.

1) Calculate the CDF F_Y of Y using

$$F_Y(y) = P(g(X) \leq y) = \int_{\{x : g(x) \leq y\}} f_X(x)\, dx.$$

2) Differentiate the CDF to obtain the PDF of Y:

$$f_Y(y) = \frac{d}{dy} F_Y(y).$$

Linear Function of a RV:

$$Y = aX + b, \quad a \neq 0.$$

Discrete case –

$$P_Y(y) = P_X\left(\frac{y-b}{a}\right), \quad \text{for all } y.$$

Continuous case –

$$Y = aX + b, \quad a \neq 0$$

$$\Rightarrow \quad f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Proof. $a > 0$ (proved earlier)

Let $a < 0$.

$$P(aX + b \leq y) = P\left(X \geq \frac{y-b}{a}\right)$$

$$= 1 - P\left(X < \frac{y-b}{a}\right)$$

$$= 1 - F_X\left(\frac{y-b}{a}\right)$$

$$\Rightarrow f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Example. X is uniform on $[0, 1]$ and let $Y = \sqrt{X}$.

For $0 \leq y \leq 1$,

$$F_Y(y) = P(\sqrt{X} \le y)$$

$$= P(X \le y^2)$$

$$= y^2$$

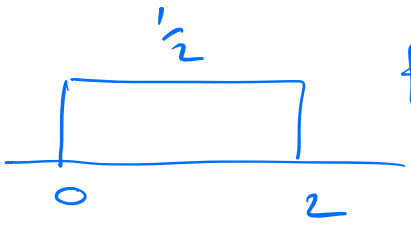
$$\Rightarrow f_Y(y) = 2y, \quad 0 \le y \le 1.$$

$F_Y(y) = 0$ $y \le 0$ and $F_Y(y) = 1$ $y \ge 1$.

$\therefore f_Y(y) = 0$ for y outside $[0, 1]$.

Example. $Y = X^3$ — $X \sim$ uniform $[0, 2]$.

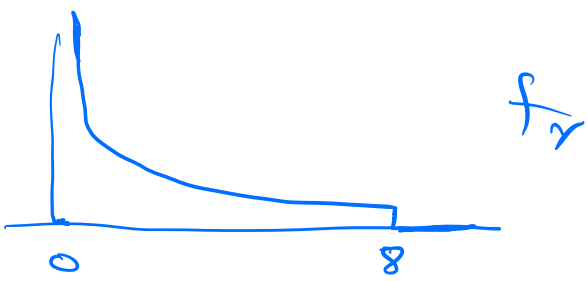
For $0 \le y \le 8$



$$F_Y(y) = P(X^3 \le y)$$

$$= P(X \le y^{1/3})$$

$$= \frac{y^{1/3}}{2}$$



$$\Rightarrow f_Y(y) = \frac{1}{6} y^{-2/3} \quad , \quad 0 \le y \le 8.$$

Example. $Y = \frac{10}{X}$, $f_X(x) = \frac{1}{5}$, $5 \le x \le 10$.

For $1 \le y \le 2$

$$F_Y(y) = P\left(\frac{10}{X} \le y\right) = P\left(X \ge \frac{10}{y}\right)$$

$$= 1 - P\left(X \le \frac{10}{y}\right)$$

$$= 1 - \frac{1}{5}\left(-5 + \frac{10}{y}\right) = -\frac{2}{y} + 2$$

$$\Rightarrow f_Y(y) = \frac{+2}{y^2}, \quad 1 \le y \le 2.$$

Monotonic Function:

Let X be a continuous RV and suppose that its range is contained in a certain interval I, in the sense that $f_X(x) = 0$ for $x \notin I$. We consider $Y = g(X)$, and assume that g is strictly monotonic over the interval I:

$g(x) < g(x')$ for all x, x' s.t. $x < x'$ (monotonically increasing)

or

$g(x) > g(x')$ for all x, x' s.t. $x < x'$. (monotonically decreasing)

For $y \in g^{-1}(I)$,

$$P(g(X) \le y)$$

$$\overset{\text{incr. } g}{=} P(X \le g^{-1}(y)) = F_X(g^{-1}(y))$$

$$\Rightarrow f_Y(y) = f_X(g^{-1}(y))\,(g^{-1})'(y).$$

For decreasing g

$$F_Y(y) = P(g(x) \le y) = P(x \ge g^{-1}(y))$$

$$= 1 - P(x \le g^{-1}(y)) = 1 - F_x(g^{-1}(y))$$

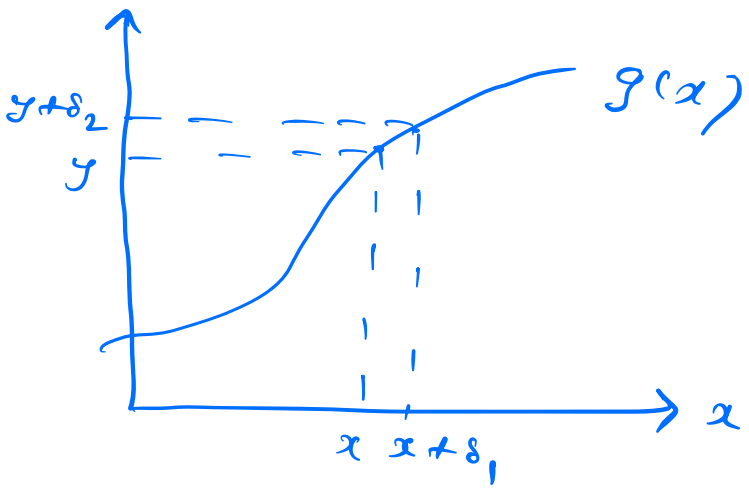
$$\Rightarrow f_Y(y) = f_x(g^{-1}(y)) \left(-(g^{-1})'(y)\right).$$

$$\therefore f_Y(y) = f_x(g^{-1}(y)) \left|(g^{-1})'(y)\right|.$$

This can also be written as

$$f_Y(y) = \begin{cases} f_x(g^{-1}(y)) / |g'(g^{-1}(y))| & \text{if } g(x) = y \text{ for some } x \text{ s.t. } f_x(x) > 0 \\ 0 & \text{if } g(x) \neq y \text{ for all } x \text{ with } f_x(x) > 0. \end{cases}$$

Illustration.



$\frac{\delta_2}{\delta_1} = g'(x)$

$\delta_1, \delta_2 \approx 0$

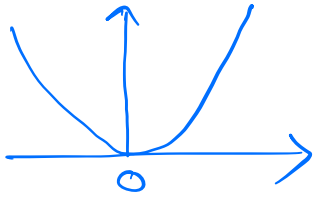
$$P(y \le Y \le y+\delta_2) \approx f_Y(y)\delta_2$$

$$= P(x \le X \le x+\delta_1) = f_x(x)\delta_1$$

$$f_Y(y) = f_X(x)\frac{\delta_1}{\delta_2} = \frac{f_X(x)}{g'(x)} = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))}.$$

Non-monotonic function

$Y = g(X)$ $\quad g(x) = x^2$



For $y \geq 0$

$$P(g(x) \leq y) = P(x^2 \leq y)$$

$$= P(-\sqrt{y} \leq x \leq \sqrt{y})$$

$$= F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

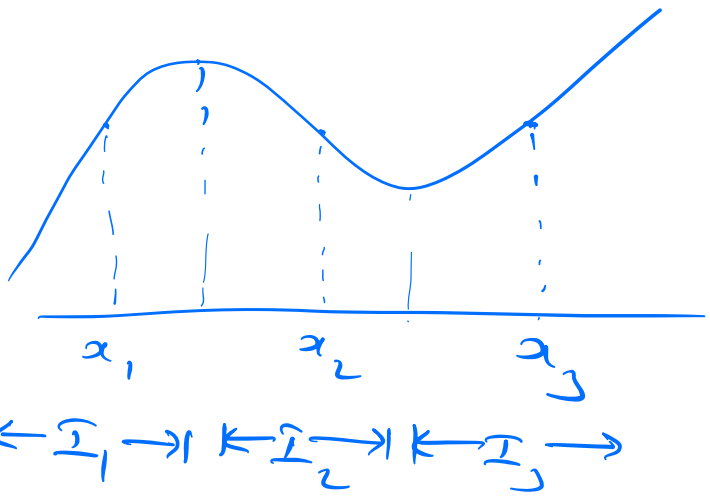
$$\Rightarrow f_Y(y) = \frac{f_X(\sqrt{y})}{2\sqrt{y}} + \frac{f_X(-\sqrt{y})}{2\sqrt{y}}.$$

Theorem. Consider a continuous RV X with PDF f_X and let $Y = g(X)$. Suppose we partition the domain of f_X into a finite number of intervals such that $g(x)$ is strictly monotone and differentiable on each interval. Then the PDF of Y is given by

$$f_y(y) = \sum_{i=1}^{n} \frac{f_x(x_i)}{|g'(x_i)|} \quad \text{where}$$

$x_1, x_2, \ldots, x_n$ are real solutions to $g(x) = y$ in the domain of f_x.

Proof.



$I_1 = [l_1, u_1]$

$I_2 = [l_2, u_2]$

$I_3 = [l_3, u_3]$.

For a given y let $g(x_i) = y$ $i \in \{1, 2, 3\}$.

$g(x)$ is increasing at x_1

decreasing at x_2 and

increasing at x_3.

$$P(g(x) \le y) = P(\{l_1 \le x \le x_1\} \cup \{x_2 \le x \le u_2\} \cup \{l_3 \le x \le x_3\})$$

$$= P(l_1 \le x \le x_1) + P(x_2 \le x \le u_2) + P(l_3 \le x \le x_3)$$

$$\Rightarrow f_y(y) = f_x(x_1)\frac{dx_1}{dy} - f_x(x_2)\frac{dx_2}{dy} + f_x(x_3)\frac{dx_3}{dy}$$

$$= \frac{f_x(x_1)}{|g'(x_1)|} + \frac{f_x(x_2)}{|g'(x_2)|} + \frac{f_x(x_3)}{|g'(x_3)|}.$$

→ For $y = x^2$, $x_1 = \sqrt{y}$, $x_2 = -\sqrt{y}$.

So $f_y(y) = \frac{f_x(\sqrt{y})}{2\sqrt{y}} + \frac{f_x(-\sqrt{y})}{2\sqrt{y}}$.

Lecture 19

(26 October 2023)

Functions of Two Random Variables

$$Z = g(X, Y),$$

i.e., $Z(\omega) = g(X(\omega), Y(\omega))$.

Example. $Z = \max(X, Y)$, X & Y are independent.

$$P(Z \le z) = P(\max(X, Y) \le z)$$

$$= P(X \le z, Y \le z)$$

$$= P(X \le z, Y \le z)$$

$$= F_X(z) F_Y(z)$$

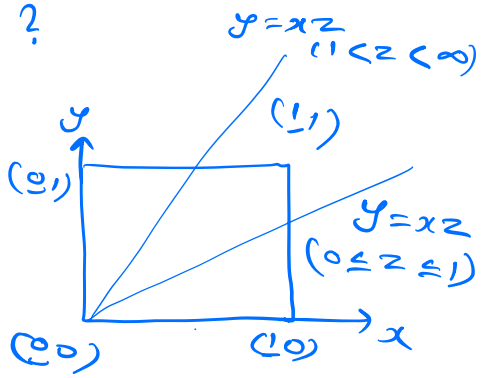
$$\Rightarrow f_Z(z) = F_X(z) f_Y(z) + f_X(z) F_Y(z).$$

As a special case $X, Y \sim$ Uniform $[0, 1]$.

$$f_Z(z) = \begin{cases} 2z, & 0 \le z \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

Exercise: $Z = \min\{X, Y\}$, X and Y are indep. Find f_Z.

Example. X and Y are independent RVs that are uniformly distributed on $[0,1]$. What is the PDF of $Z = Y/X$?



$$P(Z \le z) = P(Y/X \le z)$$

For $0 \le z \le 1$

$$P(Y \le Xz) = \int_{x=0}^{1} \int_{y=0}^{zx} 1 \, dx \, dy = z/2.$$

For $1 < z < \infty$

$$P(Y \le Xz) = 1 - \frac{1}{2z}.$$

$$\Rightarrow f_Z(z) = \begin{cases} z/2 & \text{if } 0 \le z \le 1 \\ 1 - 1/2z & \text{if } z > 1 \\ 0 & \text{o.w.} \end{cases}$$

$$f_Z(z) = \begin{cases} 1/2 & \text{if } 0 \le z \le 1 \\ 1/2z^2 & \text{if } z > 1 \\ 0 & \text{o.w.} \end{cases}$$

Sum of Independent Random Variables

Let $Z = X + Y$ where X and Y are independent random variables.

Discrete case:

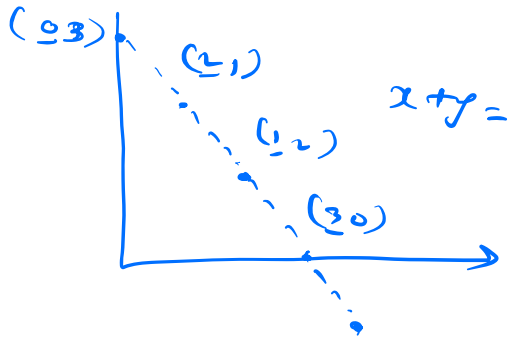
$$P_Z(z) = P(X+Y=z)$$

$$= \sum_{(x,y):\, x+y=z} P_{XY}(x,y)$$

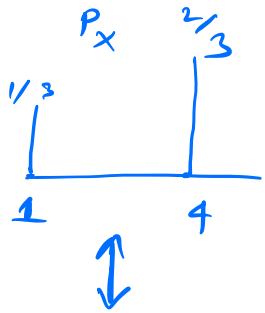
$$= \sum_x P_{XY}(x, z-x) = \sum_x P_X(x)\, P_Y(z-x).$$

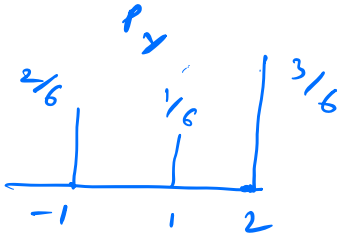
The resulting PMF P_Z is called the convolution of the PMFs of X and Y.

$Z = 3 \Rightarrow$

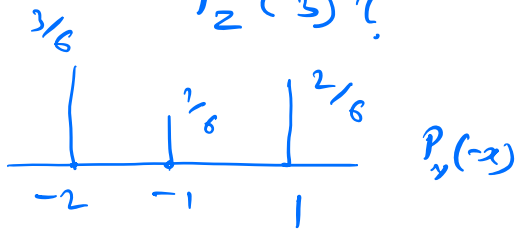


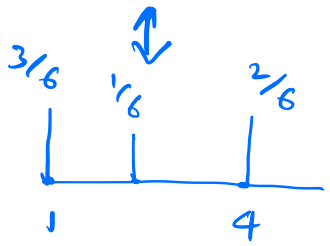
Sum of joint PMF at all points on line $x+y=3$.





$P_Z(3)$?





$P_y(3-x)$

$$\Rightarrow \sum_x P_x(x)\, P_y(3-x) = P_z(3).$$

Continuous case:

$$P(X+Y \le z) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{z-x} f_{xy}(x,y)\, dy\, dx$$

$$= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{z-x} f_x(x)\, f_y(y)\, dy\, dx$$

$$= \int_{x=-\infty}^{\infty} f_x(x)\, F_y(z-x)\, dx$$

$$\Rightarrow f_z(z) = \frac{d}{dz} F_z(z)$$

$$= \int_{-\infty}^{\infty} f_x(x)\, \frac{d}{dz} F_y(z-x)\, dx$$

$$= \int_{x=-\infty}^{\infty} f_x(x)\, f_y(z-x)\, dx .$$

Theorem. If $X \sim N(0,1)$, $Y \sim N(0,1)$ are independent then

$$X+Y \sim N(0,2).$$

Proof. $Z = X+Y$.

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

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-x)^2}{2}}\,dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2-z^2-x^2+2zx}{2}}\,dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2-zx+z^2}{2}}\,dx$$

$$= \frac{1}{\sqrt{4\pi}} \cdot e^{-z^2/4} \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-(x-z/2)^2}\,dx$$

$$= \frac{1}{\sqrt{4\pi}} \cdot e^{-z^2/4} \cdot \underbrace{\qquad\qquad}_{=1}$$

Exercise. $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$ indep $\Rightarrow X+Y \sim N(\mu_x+\mu_y, \sigma_x^2+\sigma_y^2)$.

Sum of a discrete RV and a continuous RV:

$$Z = X + Y$$

X - discrete Y - continuous X & Y are independent

$$P(Z \le z) = P(X+Y \le z)$$

$$= \sum_x P(X+Y \le z \mid X=x)\, P_X(x)$$

$$= \sum_x P(Y \le z-x \mid X=x)\, P_X(x)$$

$$= \sum_x P(Y \le z-x)\, P_X(x)$$

$$= \sum_x F_Y(z-x)\, P_X(x)$$

$$= \sum_x \int_{-\infty}^{z-x} f_Y(y)\, dy\, P_X(x) \qquad y = t-x$$

$$= \sum_x \int_{-\infty}^{z} f_Y(t-x)\, dt\, P_X(x)$$

$$= \int_{-\infty}^{z} \underbrace{\left(\sum_x f_Y(t-x)\, P_X(x) \right)}_{f_Z(t)} dt$$

$$\therefore f_Z(z) = \sum_x f_Y(z-x)\, P_X(x).$$

Functions of Two Random Variables

$(X, Y) \sim f_{XY}$.

$Z = g_1(X, Y)$, $W = g_2(X, Y)$.

$$P(Z \le z, W \le w) = P(g_1(X,Y) \le z, g_2(X,Y) \le w)$$

$$= \iint_{\substack{x,y: \\ g_1(x,y) \le z \\ g_2(x,y) \le w}} f_{XY}(x, y)\, dx\, dy$$

Sometimes it may be easier to find out the region under the above integral and compute the integral. Then we can find f_{ZW} by taking double derivative:

$$f_{ZW}(z, w) = \frac{\partial^2 F_{ZW}(z, w)}{\partial z\, \partial w}.$$

Exercise. Suppose x and y are independent uniformly distributed RVs in $[0,1)$.

$Z = \min\{x,y\}$ $W = \max\{x,y\}$. Find f_{ZW}.

- In some cases we may be able to obtain a closed-form expression for f_{ZW} in terms of f_{xy}.

Assume that g_1 & g_2 satisfy:

(1) $z = g_1(x,y)$, $w = g_2(x,y)$

$\Leftrightarrow$ $x = h_1(z,w)$ $y = h_2(z,w)$.

Example: $z = x+y$ $w = x-y$.

(2) The functions g_1 & g_2 have continuous partial derivatives at all points (x,y) and are such that

$$J(x,y) = \begin{vmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{vmatrix} = \frac{\partial g_1}{\partial x}\frac{\partial g_2}{\partial y} - \frac{\partial g_1}{\partial y}\frac{\partial g_2}{\partial x} \neq 0$$

at all points (x,y).

Theorem. In the setting above,

$$f_{ZW}(z,w) = f_{XY}(x,y)\,|J(x,y)|^{-1}.$$

This is analogous to

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|},$$

when $Y = g(X)$ for a monotonic g.

Lecture 20

(30 October 2023)

Two Functions of Two Random Variables

$(X, Y) \sim f_{XY}$.

Let $Z = g_1(X, Y)$, $W = g_2(X, Y)$.

Suppose g_1 & g_2 are such that

(i) $\exists$ functions h_1, h_2 satisfying

$$x = h_1(Z, W), \quad y = h_2(Z, W).$$

(ii)
$$J(x, y) := \begin{vmatrix} \frac{\partial g_1(x,y)}{\partial x} & \frac{\partial g_1(x,y)}{\partial y} \\ \frac{\partial g_2(x,y)}{\partial x} & \frac{\partial g_2(x,y)}{\partial g_2} \end{vmatrix}$$

$$= \frac{\partial g_1}{\partial x}\frac{\partial g_2}{\partial y} - \frac{\partial g_2}{\partial x}\frac{\partial g_1}{\partial y} \neq 0$$

at all x, y.

Theorem. In the setting above,

$$f_{ZW}(z,w) = f_{XY}(x,y)\,|J(x,y)|^{-1},$$

where $x = h_1(z,w)$ $y = h_2(z,w)$.

This is analogous to

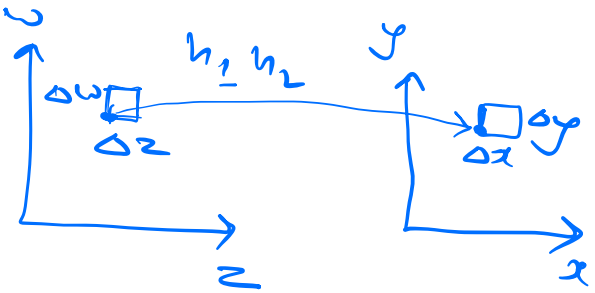
$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|},$$

when $Y = g(X)$ for a monotonic g.

Proof. Consider

$$P(z \le Z \le z+\delta z,\ w \le W \le w+\delta w)$$

$$= f_{ZW}(z,w)\,\delta z\,\delta w.$$



Alternately,

$$P(z \le Z \le z+\delta z,\ w \le W \le w+\delta w)$$

$$= P(x \le X \le x+\delta x,\ y \le Y \le y+\delta y) = f_{XY}(x,y)\,\delta x\,\delta y$$

where $x = h_1(z,w)$ $y = h_2(z,w)$.

$$\Rightarrow f_{ZW}(z,w) = f_{XY}(x,y)\,\frac{\Delta x\,\Delta y}{\Delta z\,\Delta w}$$

$$= \frac{f_{XY}(x,y)}{\left(\frac{\Delta z\,\Delta w}{\Delta x\,\Delta y}\right)} = \frac{f_{XY}(x,y)}{|J(x,y)|},$$

where

$$J(x,y) = \frac{\partial g_1}{\partial x}\frac{\partial g_2}{\partial y} - \frac{\partial g_1}{\partial y}\frac{\partial g_2}{\partial x}.$$

[See chapter 6 from Papoulis & Pillai's Book for more details and a formal justification]

Analogy: $y = g(x) \Rightarrow \frac{\Delta y}{\Delta x} = g'(x)$,

Similarly $z = g_1(x,y)$ $w = g_2(x,y)$

$$\Rightarrow \frac{\Delta z\,\Delta w}{\Delta x\,\Delta y} = \left|\det\begin{bmatrix} \partial g_1/\partial x & \partial g_1/\partial y \\ \partial g_2/\partial x & \partial g_2/\partial y \end{bmatrix}\right|.$$

- Note that Jacobian is essentially doing the following change of variables:

$$P((Z,W)\in B) = P((X,Y)\in A) = \iint_A f_{XY}(x,y)\,dx\,dy$$

$$= \iint_B f_{XY}(h_1(z,w), h_2(z,w))\cdot \frac{1}{|J(h_1(z,w), h_2(z,w)|}\,dz\,dw,$$

→ more generally for a given point (z, w) $g_1(x,y) = z$ $g_2(x,y) = w$ can have many solutions.

Let $(x_1,y_1)\ (x_2,y_2)\ \cdots\ (x_n,y_n)$ represent these multiple solutions such that

$$g_1(x_i,y_i) = z \quad g_2(x_i,y_i) = w \quad i \in [1:n].$$

Then

$$f_{zw}(z,w) = \sum_{i=1}^{n} f_{xy}(x_i,y_i)\,|J(x_i,y_i)|^{-1}.$$

Example. $\begin{bmatrix} z \\ w \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$.

$$J(x,y) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} \neq 0.$$

$$f_{zw}(z,w) = \frac{f_{xy}(x,y)}{|a_{11}a_{22} - a_{12}a_{21}|}, \quad \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^{-1} \begin{bmatrix} z \\ w \end{bmatrix}.$$

Moment Generating Functions

— A transform associated with a random variable and provides an alternative representation of a probability law.

— Particularly useful for certain types of mathematical manipulations

— Moment generating function (MGF) of a RV X is a function $M_X(s)$ defined as

$$M_X(s) = E[e^{sX}], \quad \text{for a scalar } s.$$

Discrete case:

$$M_X(s) = \sum_x e^{sx} P_X(x).$$

Continuous case:

$$M_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\, dx.$$

$\nearrow \; M_X(0) = 1.$

Example. Let $P_X(1) = \frac{1}{2}$, $P_X(2) = \frac{1}{4}$, $P_X(3) = \frac{1}{4}$.

$$M_X(s) = \frac{1}{2}e^{s} + \frac{1}{4}e^{2s} + \frac{1}{4}e^{3s}.$$

Example. Exponential RV with parameter λ.

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

$$M_X(s) = \int_0^\infty e^{sx}\,\lambda e^{-\lambda x}\,dx$$

$$= \lambda \int_0^\infty e^{(s-\lambda)x}\,dx$$

$$= \lambda \left. \frac{e^{(s-\lambda)x}}{(s-\lambda)} \right|_0^\infty$$

$$= \frac{\lambda}{\lambda - s} \quad \text{if} \quad s < \lambda,$$

otherwise the integral is infinite.

— $M_X(s)$ is only defined for those values of s for which $E[e^{sx}]$ is finite.

Exercise.

1) If X takes only non-negative integer values then show that

$$\lim_{s \to -\infty} m_X(s) = P(X=0).$$

2) Let X be a Poisson RV with parameter λ. Find $m_X(s)$.

3) Prove that $m_{aX+b}(s) = e^{sb} m_X(as)$.

Example. $X \sim N(0,1)$.

$$m_X(s) = \int_{-\infty}^{\infty} e^{sx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2} + sx} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-s)^2}{2}} \cdot e^{s^2/2} dx$$

$$= e^{s^2/2} \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-s)^2/2} dx$$

$$= e^{s^2/2}.$$

$Y \sim N(\mu, \sigma^2) \Rightarrow M_Y(s) = e^{sb} M_X(as)$

$[Y = \sigma X + \mu]$

$= e^{sb} \cdot e^{\sigma^2 s^2/2}$

$= e^{\mu s + \sigma^2 s^2/2}$.

- MGFs are useful mainly for two reasons

(1) MGFs of x gives us all moments of x.

$$M_x(s) = E[e^{sx}]$$

$$= \int_{-\infty}^{\infty} e^{sx} f_x(x)\, dx$$

$$\frac{d}{ds} M_x(s) = \int_{-\infty}^{\infty} x\, e^{sx} f_x(x)\, dx$$

$$\frac{d}{ds} M_x(s)\Big|_{s=0} = \int_{-\infty}^{\infty} x f_x(x)\, dx$$

Similarly $\frac{d^n}{ds^n} M_x(s)\Big|_{s=0} = \int_{-\infty}^{\infty} x^n f_x(x)\, dx.$

(2) MGF (if it exists) uniquely determines

the CDF of random variable.

Consider two random variables X and Y. Suppose that there exists a positive constant c such that MGFs of X and Y are finite and identical for all values of s in $[-c, c]$. Then, ⟸ Uniqueness property

$$F_X(t) = F_Y(t) \text{ for all } t \in R.$$

proof follows from connection with Laplace transform.

This can also be seen through characteristic functions.

$$\Phi_X(t) = E[e^{itx}] \quad i = \sqrt{-1}.$$

- characteristic functions always exist.
- Related to fourier transform.

Fact:

$\int_{-\infty}^{\infty} g(x)\,dx$ exists if $\int_{-\infty}^{\infty} |g(x)|\,dx$ exists.

since $|e^{itx}| = 1$ Φ_X always exists.

— If X and Y are independent RVs

$$M_{X+Y}(s) = M_X(s)\, M_Y(s).$$

Exercise. (i) $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$

$\Rightarrow X+Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

(ii) Let X and Y be independent Poisson RVs with parameters λ_1 & λ_2, respectively. Then $X+Y$ is Poisson with parameter $\lambda_1 + \lambda_2$

Lecture 21

(2 November 2023)

Module 5 : Probability Bounds & Limit Theorems

- Suppose we want to compute $P(X \geq a)$. In some scenarios it may be sufficient to have bounds on this probability instead of its exact value, e.g., when the distribution of X is unavailable or hard to compute. In such scenarios if we have exact values or bounds for the mean and variance of X, we can obtain meaningful bounds on the quantity of interest.

Markov's Inequality

If X is a non-negative random variable then $P(X \geq a) \leq \frac{E[X]}{a}$, for all $a > 0$.

Interpretation:

"If $x \geq 0$ and $E[X]$ is small, then the probability that x takes a large value must be small".

$E[X] = \sum_x x P_X(x)$, $P_X(100) \to$ small

proof. we first prove that if $x \geq y$, then $E[X] \geq E[Y]$.

$z = x - y \geq 0$, i.e., $x(\omega) - y(\omega) \geq 0$, $\forall \omega \in \Omega$.

$E[z] \geq 0 \Rightarrow E[X] \geq E[Y]$.

Let $Y = a\,1\{X \geq a\}$. we have

$X \geq Y$.

$\Rightarrow E[X] \geq E[Y]$

$\Rightarrow E[X] \geq E[a\,1\{X \geq a\}]$

$= a\,P(X \geq a)$

$\therefore P(X \geq a) \leq \frac{E[X]}{a}$.

Example. Let $X \sim \text{Binomial}(n, p)$. Using Markov's inequality find an upper bound on $P(X \geq \alpha n)$ where $p < \alpha < 1$. Evaluate the bound for $p = 1/2$ and $\alpha = 3/4$.

$$P(X \geq \alpha n) \leq \frac{E[X]}{\alpha n} = \frac{np}{n\alpha} = p/\alpha$$

$$= 1/2 \cdot \frac{4}{3} = 2/3.$$

$$\therefore P(X \geq \alpha n) \leq \frac{2}{3}.$$

Example. $X \sim \text{Uniform}[-4, 4]$.

$$P(X \geq 3) \leq P(|X| \geq 3)$$

$$\leq \frac{E[|X|]}{3}$$

$$= \frac{2}{3}.$$

Chebyshev's Inequality

If X is a random variable with mean μ and variance σ^2 then

$$P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2} \quad \text{for all } c>0.$$

Proof. Let $Y = |X-\mu|$.

$$P(Y \geq c) = P(Y^2 \geq c^2)$$

$$\leq \frac{E[Y^2]}{c^2} = \frac{E[|X-\mu|^2]}{c^2}$$

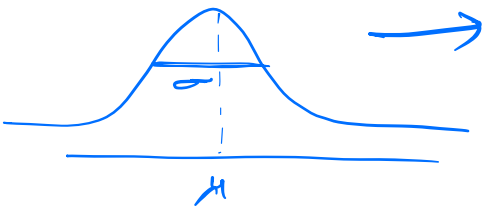
$$= \frac{\sigma_X^2}{c^2}.$$

Interpretation:

"If a random variable has small variance then the probability that it takes a value far from its mean is also small."

- Recall that variance measures the spread

of RV x around its mean.



→ small σ
⇒ dispersion around mean is small

- An alternative form of chebyshev's inequality:

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Example. $X \sim$ Binomial (n, p). Using chebyshev's inequality find an upper bound on $P(X \geq \alpha n)$ where $p < \alpha < 1$. Evaluate for $p = \frac{1}{2}$, $\alpha = \frac{3}{4}$.

$$P(X \geq \alpha n) = P(X - np \geq n(\alpha - p))$$

$$\leq P(|X - np| \geq n(\alpha - p))$$

$$\leq \frac{\text{Var}(X)}{n^2(\alpha-p)^2} = \frac{np(1-p)}{n^2(\alpha-p)^2}$$

$$= \frac{4}{n}.$$

Chernoff Bounds

If X is a random variable then for any $a \in \mathbb{R}$ we can write

$$P(X \geq a) = P(e^{sX} \geq e^{sa}) \quad \text{for } s > 0$$

$$P(X \leq a) = P(e^{sX} \geq e^{sa}) \quad \text{for } s < 0.$$

Note that e^{sX} is always a positive random variable for all $s \in \mathbb{R}$. Thus we can apply Markov's inequality.

For $s > 0$ $\quad P(X \geq a) = P(e^{sX} \geq e^{sa})$

$$\leq \frac{E[e^{sX}]}{e^{sa}}$$

$$\leq \min_{s>0} E[e^{sX}] \cdot e^{-as}.$$

Similarly for $s < 0$

$$P(X \leq a) \leq \min_{s<0} E[e^{sX}] e^{-as}.$$

Example. $X \sim \text{Binomial}(n, p)$. Using chernoff bounds bound $P(X \geq \alpha n)$ where $p < \alpha < 1$. Evaluate the bound for $p = \frac{1}{2}$ and $\alpha = \frac{3}{4}$.

$$X = \sum_{i=1}^{n} X_i \qquad P_{X_i}(1) = p = 1 - P_{X_i}(0)$$

$X_i \sim$ i.i.d.

$$E[e^{sX}] = M_X(s)$$

$$= \prod_{i=1}^{n} M_{X_i}(s)$$

$$= (pe^s + 1-p)^n.$$

$$P(X \geq \alpha n) \leq \min_{s>0} e^{-\alpha n s} (pe^s + 1-p)^n$$

$$\frac{d}{ds}\left(e^{-\alpha n s}(pe^s + 1-p)^n\right) = 0$$

$$\Rightarrow e^s = \frac{\alpha(1-p)}{p(1-\alpha)}$$

$$\Rightarrow s = \log \frac{\alpha(1-p)}{p(1-\alpha)} > 0$$

Since $\frac{1-p}{p} \cdot \frac{\alpha}{1-\alpha} > 1$,

Also check double derivative ≥ 0.

we get

$$P(X \geq \alpha n) \leq \left(\frac{\alpha(1-p)}{p(1-\alpha)}\right)^{-\alpha n} \left(\frac{p\alpha(1-p)}{p(1-\alpha)} + 1-p\right)^{n}$$

$$= \left(\frac{\alpha(1-p)}{p(1-\alpha)}\right)^{-\alpha n} \left(\frac{1-p}{1-\alpha}\right)^{n}$$

$$= \left(\frac{1-p}{1-\alpha}\right)^{(1-\alpha)n} \left(\frac{p}{\alpha}\right)^{\alpha n}$$

$$= \left(\frac{1-1/2}{1-3/4}\right)^{\frac{n}{4}} \cdot \left(\frac{2}{3}\right)^{3n/4}$$

$$= \frac{2^{n/4} \cdot 2^{3n/4}}{27^{n/4}} = \left(\frac{16}{27}\right)^{n/4}.$$

comparison between markov, chebyshev and chemoff bounds:

$P(X \geq \alpha n) \leq \frac{2}{3}$ [markov]

$P(X \geq \alpha n) \leq 4/n$ [chebyshev]

$P(X \geq \alpha n) \leq \left(\frac{16}{27}\right)^{n/4}$ [chemoff]

- The bound given by Markov is the weakest bound. It is a constant (does not depend on n).
- Chebyshev's bound is stronger than Markov's. In particular note that $4/n \to 0$ as $n \to \infty$.
- Chernoff bound is the strongest bound. It goes to zero exponentially fast.

Example. Suppose X is a RV taking values in $[a,b]$. Obtain a bound on $P(|X-\mu| \geq c)$ using Chebyshev's inequality.

claim: $X \in [a,b] \Rightarrow \sigma_X^2 \leq \frac{(b-a)^2}{4}$.

$E[(X-v)^2] = E[X^2] - 2E[X]v + v^2$

$\sigma_X^2 \leq E[(X - \frac{a+b}{2})^2] = E[((X-a)+(X-b))^2]/4$

$= \frac{1}{4} E[((X-a)-(X-b))^2 + 4(X-a)(X-b)]$

$\leq (b-a)^2/4$.

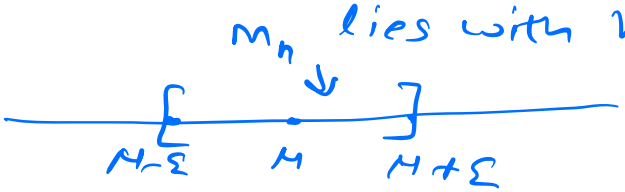
$P(|X-\mu| \geq c) \leq \frac{\sigma_X^2}{c^2} \leq \frac{(b-a)^2}{4c^2}$.

Weak Law of Large Numbers (WLLN)

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean μ. For every $\varepsilon > 0$, we have

$$P\left(\left| \frac{\sum_{i=1}^{n} X_i}{n} - \mu \right| \geq \varepsilon \right) \to 0 \quad \text{as} \quad n \to \infty.$$

Interpretation: Let $M_n = \sum_{i=1}^{n} X_i / n$. WLLN asserts that the sample mean of a large number of i.i.d. RVs is very close to the true mean.



Proof. $E[M_n] = \mu$, $\text{Var}(M_n) = \frac{\sigma^2}{n}$.

$$P(|M_n - E[M_n]| \geq \epsilon)$$

$$= P(|M_n - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(M_n)}{\epsilon^2}$$

$$= \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0 \quad \text{as} \quad n \to \infty,$$

Lecture 22

(6 November 2023)

WLLN. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed RVs with mean μ. For every $\varepsilon > 0$ we have

$$P\left(\left| \frac{\sum_{i=1}^{n} X_i}{n} - \mu \right| \geq \varepsilon \right) \to 0 \quad \text{as} \quad n \to \infty.$$

Remark. Note that the proof we have seen assumes finite variance. It turns out that this law remains true even if the X_i have infinite variance, but a much more elaborate argument is needed.

Another Interpretation:

Let $a_1, a_2, \ldots, a_n$ be the realizations of a sequence of i.i.d. random variables $\sim p_X$.

$$a_i \in \mathcal{X} = \{x_1, x_2, \ldots, x_k\}.$$

$$\sum_{i=1}^{n} a_i / n = \sum_{i=1}^{k} x_i k_i / n \quad , \quad \sum_{i=1}^{k} k_i = n$$

k_i = no. of times x_i occurs

If we interpret probabilities as relative frequency

$$\sum_{i=1}^{k} \left(\frac{k_i}{n}\right) x_i \longrightarrow \sum_{i=1}^{k} p_X(x_i)\, x_i = E[X].$$

Example (Polling). Each voter in a population selects a particular candidate A with probability p and chooses another with probability $1-p$.

$$X_i = \begin{cases} 1 & i^{th} \text{ voter chooses A} \\ 0 & \text{o.w.} \end{cases}$$

$$p_{X_i}(1) = p = 1 - p_{X_i}(0).$$

X_i are independent and identically distributed.

We are interested in knowing the value of p.

$$M_n = \sum_{i=1}^{n} X_i / n.$$

$$P(|M_n - p| \geq 0.1) \leq \frac{p(1-p)}{n(0.1)^2} \leq \frac{1}{400(0.1)^2} \quad \text{for } n = 100$$

$$= 0.25$$

With a sample size of 100 the probability that our estimate is incorrect by more than 0.1 is smaller than 0.25.

$$P(|M_n - p| \geq 0.01) \leq \frac{1}{4n(0.01)^2}$$

$$n \geq 50000 \Rightarrow P(|M_n - p| \geq 0.01) \leq 0.05.$$

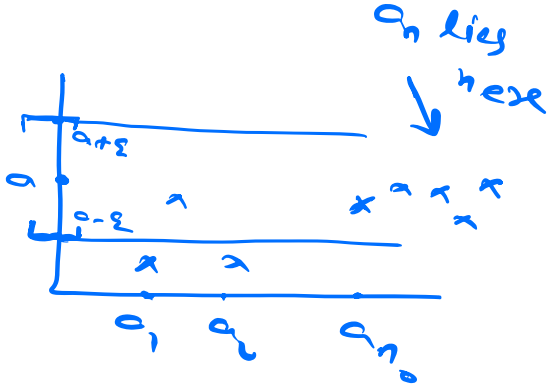
Convergence in Probability

We can interpret WLLN as stating that "$M_n = \sum_{i=1}^{n} X_i/n$ converges to mean M" in the sense of "convergence in probability".

Recall the convergence of a deterministic sequence.

Let $a_1, a_2, \ldots$ be a sequence of real numbers and let a be another real. We say that a_n converges to a or $\lim_{n\to\infty} a_n = a$ if for every $\epsilon > 0$ there exists some n_0 s.t.

$|a_n - a| \leq \varepsilon$ for all $n \geq n_0$.



Convergence in Probability

Let $Y_1, Y_2, \ldots$ be a sequence of random variables, and let a be a real number. We say that Y_n converges to a in probability if for every $\varepsilon > 0$ we have

$$\lim_{n \to \infty} P(|Y_n - a| \geq \varepsilon) = 0.$$

- If RVs $Y_1, Y_2, \ldots$ have a PMF or a PDF and converge in probability to a, then "almost all" of the PMF or PDF of Y_n is concentrated within ε of a for large values of n.
- The above condition of convergence in probability can be equivalently written as follows: for every $\varepsilon > 0$ and for every $\delta > 0$ there exists some n_0 such that

$$P(|Y_n - a| \geq \varepsilon) \leq \delta \quad \text{for all } n \geq n_0.$$

$\varepsilon \rightarrow$ accuracy level, $\delta \rightarrow$ confidence level.

Example. Let $X_1, X_2, \ldots$ be a sequence of independent and uniformly distributed RVs in $[0,1]$, and let $Y_n = \min\{X_1, \ldots, X_n\}$.

Note that $Y_1 \geq Y_2 \geq \ldots$. In fact,

$$Y_n \to 0 \quad \text{in probability.}$$

$$P(|Y_n - 0| \geq \varepsilon) = P(X_1 \geq \varepsilon, \ldots, X_n \geq \varepsilon)$$

$$= (1-\varepsilon)^n \to 0 \quad \text{as } n \to \infty$$

for every $\varepsilon > 0$.

- The following example shows that if $Y_n \to a$, then $E[Y_n]$ may not converge to a.

Example. $$p(Y_n = y) = \begin{cases} 1 - 1/n, & \text{for } y = 0 \\ 1/n, & \text{for } y = n^2 \\ 0, & \text{elsewhere.} \end{cases}$$

For every $\varepsilon > 0$ $\quad P(|Y_n - 0| \geq \varepsilon) = \frac{1}{n} \to 0$ as $n \to \infty$.

$E[Y_n] = n \to \infty$ as $n \to \infty$,

i.e. $E[Y_n] \not\to a$.

Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed RVs with common mean μ and variance σ^2.

Define $Z_n = \dfrac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma}$. Then

$$\lim_{n\to\infty} P(Z_n \le z) = \Phi(z)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx.$$

- $S_n = \sum_{i=1}^{n} X_i$, $\quad \mathrm{var}(S_n) = n\sigma^2 \to \infty$ as $n \to \infty$

$M_n = \sum_{i=1}^{n} X_i / n$, $\quad \mathrm{var}(M_n) = \dfrac{\sigma^2}{n} \to 0$ as $n \to \infty$

$Z_n = \sum_{i=1}^{n} X_i / \sqrt{n}$, $\quad \mathrm{var}(Z_n) = 1$ independent of n (constant)

Proof of CLT:

Fact. If $M_{Z_n}(t) \to M_Z(t)$, then

$$F_{Z_n}(z) \to F_Z(z) \text{ for all } z.$$

Proof of the fact is related to the continuity of inverse Fourier transform.

We assume that $M_X(s)$ is finite when $-d < s < d$. Let X_i has zero mean & variance 1.

$$Z_n = \sum_{i=1}^{n} X_i / \sqrt{n}.$$

$$M_{Z_n}(s) = E\left[e^{s \sum_{i=1}^{n} X_i/\sqrt{n}} \right]$$

$$= \prod_{i=1}^{n} E\left[e^{s X_i/\sqrt{n}} \right] = M_X(s/\sqrt{n})^n.$$

Let $L(t) = \log M_X(t)$.

$\mathcal{L}(0) = 0, \quad \mathcal{L}'(0) = \dfrac{M_x'(0)}{M_x(0)} = E[X] = 0.$

$\mathcal{L}''(0) = \dfrac{M_x(0)M_x''(0) - M_x'(0)^2}{M_x(0)^2} = E[X^2] = 1.$

We first show that

$$M_x(s/\sqrt{n})^n \longrightarrow e^{t^2/2},$$

↑ MGF of $N(0,1)$

or equivalently, that

$$n\mathcal{L}(s/\sqrt{n}) \longrightarrow t^2/2 \quad \text{as} \quad n \to \infty.$$

$$\lim_{n\to\infty} \frac{\mathcal{L}(s/\sqrt{n})}{n^{-1}} = \lim_{n\to\infty} \frac{-\mathcal{L}'(s/\sqrt{n}) \cdot s\, n^{-3/2}}{-2\, n^{-2}}$$

(by L'Hopital Rule)

$$= \lim_{n\to\infty} \frac{\mathcal{L}'(s/\sqrt{n}) \cdot s}{2n^{-1/2}}$$

$$= \lim_{n\to\infty} \frac{-\mathcal{L}''(s/\sqrt{n})\, s^2\, n^{-3/2}}{-2n^{-3/2}} = \lim_{n\to\infty} \mathcal{L}''(s/\sqrt{n}) \frac{s^2}{2} = \frac{s^2}{2},$$

(by L'Hopital rule) ($\because \mathcal{L}''(0) = 1$)

If x_i's are of mean μ & variance σ^2,

$$\frac{\sum_{i=1}^{n} x_i - n\mu}{\sqrt{n}\sigma} = \frac{\sum_{i=1}^{n} \frac{x_i - \mu}{\sigma}}{\sqrt{n}} \longrightarrow \begin{matrix} \text{mean } 0 \\ \text{var} = 1. \end{matrix}$$

Fact. If $M_{Z_n}(t) \to M_Z(t)$, then

$$F_{Z_n}(z) \to F_Z(z) \text{ for all } z.$$

Proof is related inverse Fourier transform and its continuity.

This completes the proof of the central limit theorem.

Lecture 23

(9 November 2023)

Central Limit Theorem.

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with common mean μ and variance σ^2. Define

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}.$$ Then, the CDF of Z_n converges to the standard normal CDF in the sense that

$$\lim_{n\to\infty} P(Z_n \le z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx,$$

for every z.

Proof. We prove this for $\mu = 0$ & $\sigma^2 = 1$, i.e., $Z_n = \sum_{i=1}^{n} X_i/\sqrt{n}$.

$$M_{Z_n}(s) = E\left[e^{sZ_n}\right]$$

$$= E\left[e^{s\sum_{i=1}^{n} x_i/\sqrt{n}}\right]$$

$$= \prod_{i=1}^{n} M_{x_i}(s/\sqrt{n}) = \left(M_x(s/\sqrt{n})\right)^n.$$

Let $L(s) = \log M_x(s)$.

$$\log M_{Z_n}(s) = n \log M_x(s/\sqrt{n}) = n L(s/\sqrt{n}).$$

We have $L(0) = L'(0) = 0$ & $L''(0) = 1$.

Consider $\lim_{n\to\infty} n \log M_x(s/\sqrt{n})$

$$= \lim_{n\to\infty} \frac{\log M_x(s/\sqrt{n})}{n^{-1}}$$

$$= \lim_{n\to\infty} \frac{M_x'(s/\sqrt{n})\, s\, n^{-3/2}}{2 M_x(s/\sqrt{n})\, n^{-2}}$$

$$= \lim_{n\to\infty} \frac{L'(s/\sqrt{n})\, s}{2\, n^{-1/2}}$$

$$= \lim_{n\to\infty} \frac{L''(s/\sqrt{n})\, s^2\, n^{-3/2}}{2\, n^{-3/2}} = s^2/2.$$

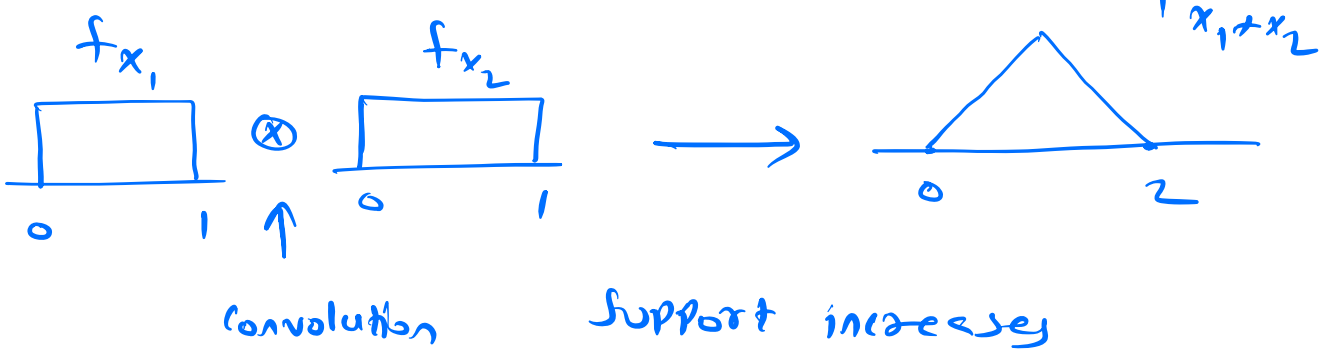
Examples. (i) $X_i \sim N(\mu, \sigma^2)$

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} \sim N(0,1) \quad \text{for all } n \in \mathbb{N},$$

$$\frac{X_1 - \mu}{\sigma}, \quad \frac{X_1 + X_2 - 2\mu}{\sqrt{2}\,\sigma}, \quad \cdots \quad N(0,1),$$

$$\lim_{n \to \infty} P(Z_n \le z) = P(Z \le z), \quad Z \sim N(0,1).$$

(ii) $X_i \sim$ uniform $[0,1]$.



Intuitively as $n \to \infty$, support will be real line and $\sum_{i=1}^{n} X_i \to$ Gaussian RV.

Normal Approximation Based on CLT:

Let $S_n = \sum_{i=1}^{n} x_i$, where x_i are i.i.d. with mean μ and variance σ^2. We are interested in computing $P(S_n \leq c)$. If n is large it can be approximated in the following way.

$$P(S_n \leq c) = P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq \frac{c - n\mu}{\sqrt{n}\sigma}\right)$$

$$\approx \Phi(z) \quad \text{where } z = \frac{c - n\mu}{\sqrt{n}\sigma}.$$

Example. X_i, $i \in [1:100]$ are i.i.d. and uniformly distributed in the interval $[5, 50]$. Find $P(S_{100} > 3000)$, $S_{100} = \sum_{i=1}^{100} x_i$. It is not easy to calculate CDF and the desired probability exactly, but an approximate answer can be quickly obtained using CLT.

$$\mu = \frac{5+50}{2}, \quad \sigma^2 = \frac{(50-5)^2}{12} = 168.75.$$

$$P(S_{100} > 3000) = 1 - P(S_{100} \leq 3000)$$

$$= 1 - P\left(\frac{S_{100} - n\mu}{\sqrt{n}\sigma} \leq \frac{3000 - 2750}{10\,(16.875)}\right)$$

$$= 1 - P\left(\frac{S_{100} - n\mu}{\sqrt{n}\sigma} \leq 1.92\right)$$

$$= 1 - \Phi(1.92)$$

$$\approx 1 - 0.9726$$

(using standard Normal table)

$$= 0.0274.$$

Example (Polling). $X_i = \begin{cases} 1 & \text{if } i^{th} \text{ voter votes for A} \\ 0 & \text{otherwise} \end{cases}$

$E[X_i] = p$ $\quad M_n = \sum_{i=1}^{n} X_i / n$.

WLLN (or Chebyshev's inequality):

For $n = 100$, we got that

$$P(|M_n - p| \geq 0.1) \leq 0.25.$$

We will see now if CLT can improve this bound.

$$P\left(\left|\frac{\sum_{i=1}^{n} x_i}{n} - p\right| \geq 0.1\right)$$

$$= P\left(\left|\frac{\sum_{i=1}^{n} x_i - np}{n}\right| \geq 0.1n\right)$$

$$= P\left(\left|\frac{\sum_{i=1}^{n} x_i - np}{\sqrt{n}\sigma}\right| \geq \frac{0.1\sqrt{n}}{\sigma}\right)$$

$$= P\left(\frac{\sum_{i=1}^{n} x_i - np}{\sqrt{n}\sigma} \geq \frac{0.1\sqrt{n}}{\sigma}\right) + P\left(\frac{\sum_{i=1}^{n} x_i - np}{\sqrt{n}\sigma} \leq \frac{-0.1\sqrt{n}}{\sigma}\right)$$

$$\approx 2P\left(Z \geq \frac{0.1\sqrt{n}}{\sigma}\right) \quad Z \sim N(0,1)$$

$$\leq 2P\left(Z \geq 0.2\sqrt{n}\right) \quad \text{as } \sigma \leq \frac{1}{2}$$

$$= 2 - 2P\left(Z \leq 0.2\sqrt{n}\right)$$

$$= 2 - 2\Phi\left(0.2\sqrt{100}\right) = 2 - 2\Phi(2) = 2 - 2 \times 0.977$$

$$= 0.046.$$

This is a much better estimate than 0.25 from WLLN.

Strong Law of Large Numbers

- SLLN also deals with the convergence of the sample mean to the true mean. However, SLLN refers to a different type of convergence.

WLLN

$$M_n = \sum_{i=1}^{n} x_i / n \longrightarrow E[x] \quad \text{in probability}$$

SLLN

$$M_n = \sum_{i=1}^{n} x_i / n \longrightarrow E[x] \quad \text{with probability 1.}$$

(or almost sure convergence)

Almost sure convergence:

Let $Y_1, Y_2, \ldots$ be a sequence of RVs and c be a real number. Then we say $Y_n \xrightarrow{a.s.} a$ if $P\left(\left\{\omega \in \Omega : \lim_{n \to \infty} x_n(\omega) = c\right\}\right) = 1$.

Example. Let $\Omega = [0,1]$. Consider a probability law defined by $P([a,b]) = b-a$ for all $0 \le a \le b \le 1$.

Define $X_n(\omega) = \omega^n$ for $n \in \mathbb{N}$.

Note that $\lim_{n\to\infty} X_n(\omega) = \begin{cases} 0, & \text{if } 0 \le \omega < 1 \\ 1, & \text{if } \omega = 1. \end{cases}$

$$P(\{\omega : \lim_{n\to\infty} X_n(\omega) = 0\}) = P([0,1)) = 1$$

Since the singleton set $\{1\}$ has zero probability,

$$\therefore X_n \xrightarrow{a.s.} 0.$$

SLLN. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. RVs with mean μ. Then,

$$M_n = \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{a.s.} \mu, \quad \text{i.e.,}$$

$$P\left(\{\omega : \lim_{n\to\infty} M_n(\omega) = \mu\}\right) = 1.$$

Proof. Assume that $E[X_i^4] = k < \infty$,

$S_n = \sum_{i=1}^{n} X_i$. $E[S_n^4] = E\left[\left(\sum_{i=1}^{n} X_i\right)^4\right]$.

This will have terms of the form

$$X_i^4, \quad X_i^3 X_j, \quad X_i^2 X_j^2, \quad X_i^2 X_j X_k, \quad X_i X_j X_k X_l$$

where i, j, k, l are different.

Assume $\mu = 0$. Then because of independence it follows that

$$E[X_i^3 X_j] = E[X_i^3] E[X_j] = 0$$

$$E[X_i^2 X_j X_k] = E[X_i^2] E[X_j] E[X_k] = 0$$

$$E[X_i X_j X_k X_l] = E[X_i] E[X_j] E[X_k] E[X_l] = 0.$$

so $E[S_n^4] = n E[X_i^4] + 6 \binom{n}{2} E[X_i^2 X_j^2]$

$$= n E[X_i^4] + 3n(n-1) \underbrace{E[X_i^2] E[X_j^2]}$$

$$= (E[X_i^2])^2 \le E[X_i^4]$$

as $\mathrm{Var}(X_i^2) \ge 0$

$$\le nk + 3n(n-1)k \le 3n^2 k$$

$$\Rightarrow E[S_n^4 / n^4] \le 3k / n^2$$

$$\Rightarrow E\left[\sum_{n=1}^{\infty} S_n^4/n^4\right] = \sum_{n=1}^{\infty} E\left[S_n^4/n^4\right]$$

$$= \sum_{n=1}^{\infty} \frac{1}{n^2} \cdot (3K) < \infty.$$

This implies that

$$P\left(\sum_{n=1}^{\infty} S_n^4/n^4 < \infty\right) = 1$$

$$\sum_{n=1}^{\infty} \frac{S_n^4}{n^4} < \infty \Rightarrow \lim_{n\to\infty} \frac{S_n^4}{n^4} = 0$$

$$\Rightarrow 1 = P\left(\sum_{n=1}^{\infty} \frac{S_n^4}{n^4} < \infty\right) \leq P\left(\lim_{n\to\infty} \frac{S_n^4}{n^4} = 0\right)$$

$$\Rightarrow P\left(\lim_{n\to\infty} \frac{S_n}{n} = 0\right) = 1.$$

In the above proof we have used

$E\left[\sum_{i=1}^{\infty} Z_i\right] = \sum_{i=1}^{\infty} E[Z_i]$. This is not necessarily true always for any Z_i's.

However this holds true when all Z_i are non-negative random variables. This is because of monotone convergence theorem (not covered in this course).

Lecture 24

(16 November 2023)

Random Processes

- A random or stochastic process is a mathematical model of a probabilistic experiment that evolves in time and generates a sequence of numerical values.
- Each numerical value in the sequence is modeled by a random variable, so a random process is simply a (finite or infinite) sequence of random variables.

Recall that a RV $X : \Omega \to \mathbb{R}$.

Formally, a random process is a family of random variables $(X_t : t \in T)$, all on the same probabilistic model (Ω, P).

In many applications, T is a set of times. If $T = \mathbb{N}$, then $(X_t : t \in T)$ is called a discrete-time process. If $T = \mathbb{R}$, then continuous-time process.

- For each $t \in T$ X_t is a RV.

$$\omega \longmapsto X_1(\omega) \quad X_2(\omega) \quad - - -$$

Discrete-time

$$\omega \longmapsto X_t(\omega) \quad t \in T.$$

Continuous-time

For a fixed $\omega \in \Omega$ $(X_t(\omega) \; t \in T)$ is called the sample path at ω.

Mean Function:

$$\mu_X(t) = E[X_t] \quad \text{for } t \in T.$$

Covariance Function:

$$C_X(t_1, t_2) = \text{Cov}(X_{t_1}, X_{t_2})$$

$$= E[X_{t_1} X_{t_2}] - E[X_{t_1}] E[X_{t_2}]$$

Correlation function:

$$R_X(t_1, t_2) = E[X_{t_1} X_{t_2}].$$

- A random process is statistically specified by its complete set of n^{th} order distribution function for all $n \in N$.

$$F_X(x_1 \dots x_n ; t_1 t_2 \dots t_n)$$

$$= P(X_{t_1} \le x_1 \dots X_{t_n} \le x_n).$$

For the continuous case we can obtain the PDF as

$$f(x_1 x_2 ; t_1 t_2) = \frac{\partial^2 F(x_1 x_2 ; t_1 t_2)}{\partial x_1 \partial x_2}.$$

Arrival-Type processes we are interested in occurances that have the character of an arrival such as message receptions at a receiver, customer purchases at a store etc. we will focus on models in which the interarrival times (the times between successive arrivals) are independent random variables

Bernoulli process - case where arrivals occur in discrete time and the interarrival times are geometrically distributed

Poisson process - case where arrivals occur in continuous time and the interarrival times are exponentially distributed.

The Bernoulli process

Consider a sequence of independent coin tosses with the probability of heads p in the range $0 < p < 1$.

- A Bernoulli process is a sequence $X_1, X_2, \ldots$ of independent Bernoulli RVs X_i with

$$P(X_i = 1) = P(\text{success at the } i\text{th trial})$$

$$P(X_i = 0) = P(\text{failure at the } i\text{th trail})$$

- No. of arrivals within a certain time period or no. of successes in n independent trails:

$$P_S(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, \ldots, n$$

$$E[S] = np, \quad \text{var}(S) = np(1-p).$$

This is Binomial with parameters n & p.

- The time until the first arrival or no. of trials up to (and including) the first success:

$$p_T(t) = (1-p)^{t-1} p, \quad t = 1, 2, \ldots$$

$$E[T] = \frac{1}{p}, \quad \text{var}(T) = \frac{1-p}{p^2}.$$

Independence & Memorylessness

Due to independence property in the Bernoulli process whatever has happened in the past trials provides no information on the outcome of future trials.

Fresh-start property:

For any given n, consider

$X_1, X_2, \ldots, X_n, X_{n+1}, X_{n+2}, \ldots$, we notice that

$Y_i = X_{n+i}$ are independent Bernoulli trials and therefore form a Bernoulli process.

Memorylessness property: T = time until 1st success

$$P(T-n=t \mid T>n)$$

$$= \frac{P(T-n=t \text{ and } T>n)}{P(T>n)}$$

$$= \frac{P(T=n+t)}{P(T>n)} = \frac{(1-p)^{n+t-1} \cdot p}{(1-p)^n}$$

$$= (1-p)^{t-1} p = P(T=t).$$

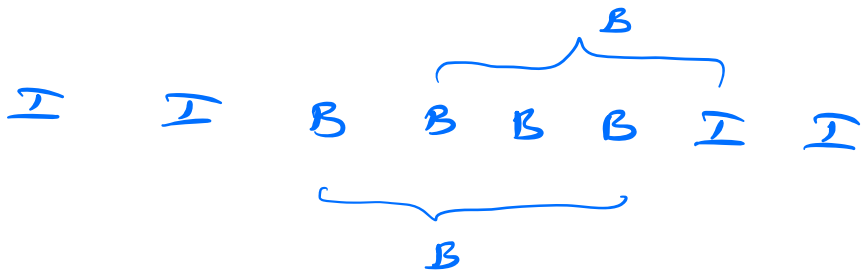
- we call a slot busy (B) if there is an arrival, and otherwise let us call it idle. we call a string of idle slots flanked by busy slots an 'idle period'. similarly we can define 'busy period'.

Let us derive the PMF, mean, and variance of the following random variables.

T = the time index of the first idle slot

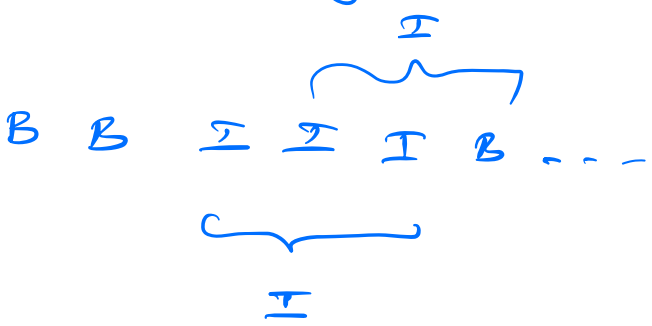
$$P_T(k) = p^{k-1}(1-p) \quad k = 1, 2, \ldots$$

L = the length (no. of slots) of the first busy period



$$P_B(k) = p^{k-1}(1-p) \quad k = 1, 2, \ldots$$

I = the length of the first idle period



$$P_I(k) = (1-p)^{k-1} p \quad k = 1, 2, \ldots$$

Example. Let N be the first time that we have a success immediately following a previous success. That is, N is the first time i for which $x_{i-1} = x_i = 1$. What is the probability that there are no successes in the two trials that follow, i.e., $P(x_{N+1} = x_{N+2} = 0)$?

$$P(X_{N+1} = X_{N+2} = 0)$$

$$= \sum_{n=1}^{\infty} P(X_{N+1} = X_{N+2} = 0 \mid N = n)\, P(N = n)$$

$$P(X_{N+1} = X_{N+2} = 0 \mid N = n)$$

$$= P(X_{n+1} = X_{n+2} = 0) = (1-p)^2$$

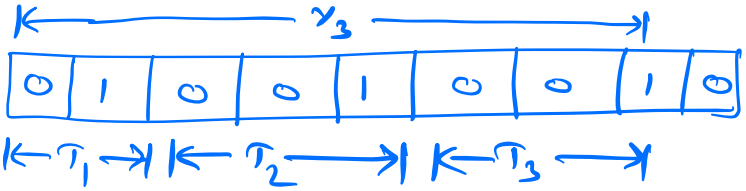
(equality because $N = n$ is determined by some function of $X_1, X_2, \ldots, X_n$).

Interarrival Times

Let Y_k denote the time of the k^{th} success (or arrival).

$$T_1 = Y_1 \qquad T_k = Y_k - Y_{k-1}, \quad k = 2, 3, \ldots$$

T_k represents the k^{th} interarrival time i.e. no. of trials following the $(k-1)^{st}$ success until the next success.



We have $Y_k = \sum_{i=1}^{k} T_i$.

Interarrival times $T_1, T_2, T_3, \ldots$ are independent and all have the same geometric distribution.

Properties of the k^{th} Arrival Time

$$Y_k = T_1 + T_2 + \cdots + T_k$$

Each T_i is a geometric RV with parameter p and all T_i's are independent.

$$E[Y_k] = \sum_{i=1}^{k} E[T_i] = k/p,$$

$$\text{Var}(Y_k) = \sum_{i=1}^{k} \text{Var}(T_i) = \frac{k(1-p)}{p^2},$$

To find the pmf of Y_k

$$P(Y_k = t) = P(A \cap B) \quad \text{where}$$

$A = \{$ trial t is a success $\}$

$B = \{$ exactly $k-1$ successes in $t-1$ trials $\}$

$$P_{Y_k}(t) = P(A) \cdot P(B)$$

$$= p \cdot \binom{t-1}{k-1} p^{k-1} (1-p)^{t-k}$$

$$= \binom{t-1}{k-1} p^{k} (1-p)^{t-k} .$$

Known as Pascal PMF of order k.

Poisson Process

- A counting process counts the number of arrivals, e.g., no. of customers who arrive at a restaurant — $X(t)$ = no. of arrivals in time t. $[X(0) = 0]$
- A counting process has independent and stationary increments.

Independent increments:

For all $0 \le t_1 < t_2 < t_3 \cdots < t_n$ the RVs

$X(t_2) - X(t_1)$, $X(t_3) - X(t_2)$, ... are independent.

Note that $X(t_i) - X(t_{i-1})$ represents the no. of arrivals in the interval $(t_{i-1}, t_i]$.

Stationary increments:

For all $t_2 > t_1 \geq 0$ & $\tau > 0$, the RVs

$X(t_2) - X(t_1)$ and $X(t_2+\tau) - X(t_1+\tau)$

have the same distribution.

- Thus a counting process has independent increments if the no. of arrivals in non-overlapping intervals are independent. Also it has stationary increments if, for all $t_2 > t_1 \geq 0$ $X(t_2) - X(t_1)$ has the same distribution as $X(t_2 - t_1)$.

- Before we define Poisson process we recall Poisson RV.

$X \sim \text{Poisson}(\lambda)$ means $P_X(k) = \frac{e^{-\lambda} \lambda^k}{k!}$, $k = 0, 1, 2, \ldots$

- A counting process $(N(t), t \in [0, \infty))$ is called a Poisson process with rate λ if:

(1) $N(0) = 0$;

(2) $N(t)$ has independent increments;

(3) the no. of arrivals in any interval of length $T > 0$ has Poisson (λT) distribution.

$E[N(t)] = \lambda t$ $\quad$ $\text{var}(N(t)) = \lambda t$.

$$R_X(t_1, t_2) = E[N(t_1) N(t_2)] \quad (t_1 < t_2)$$

$$= E[N(t_1)(N(t_2) - N(t_1) + N(t_1))]$$

$$= E[N(t_1)(N(t_2) - N(t_1))] + E[N^2(t_1)]$$

$$= E[N(t_1)] E[N(t_2 - t_1)] + E[N^2(t_1)]$$

$$= \lambda t_1 \cdot \lambda(t_2 - t_1) + \lambda t_1 + \lambda^2 t_1^2$$

$$= \lambda^2 t_1 t_2 - \lambda^2 t_1^2 + \lambda t_1 - \lambda^2 t_1^2$$

$$= \lambda t_1 + \lambda^2 t_1 t_2.$$

$$C(t_1, t_2) = \lambda \min\{t_1, t_2\}.$$

Lecture 25

(18 November 2023)

- A random process $X = (X_t : t \in T)$ is specified by n^{th} order distribution function

$$F_{X_{t_1} \cdots X_{t_n}}(x_1, x_2, \cdots, x_n)$$

$$= P(X_{t_1} \le x_1, \cdots, X_{t_n} \le x_n)$$

for all $n \in \mathbb{N}$.

Example. Consider the random process

$$X(t) = A \sin(\omega_0 t + \Theta),$$

where A and Θ are independent and Θ is uniformly distributed over $[-\pi, \pi]$.

$$\mu_X(t) = E[A \sin(\omega_0 t + \Theta)]$$

$$= \mu_A \cdot E[\sin(\omega_0 t + \Theta)]$$

$$= \mu_A \cdot \int_{-\pi}^{\pi} \frac{1}{2\pi} \cdot \sin(\omega_0 t + \theta)\, d\theta$$

$= 0.$

$$R_X(t_1, t_2) = E[X(t_1) X(t_2)]$$

$$= E[A^2 \sin(\omega_0 t_1 + \Theta) \sin(\omega_0 t_2 + \Theta)]$$

$$= E[A^2] \; E[\sin(\omega_0 t_1 + \Theta) \sin(\omega_0 t_2 + \Theta)]$$

$$= E[A^2] \frac{1}{2} \left(E[\cos(\omega_0 (t_1 - t_2))] - \underbrace{E[\cos(\omega_0 (t_1 + t_2) + 2\Theta)]}_{= 0} \right)$$

$$= \frac{1}{2} E[A^2] \; [\cos(\omega_0 (t_1 - t_2))].$$

Stationary Process

A random process $(X(t), t \in R)$ is stationary if for all $t_1, t_2, \ldots, t_n, T \in R$

$$F_{X(t_1) X(t_2) \ldots X(t_n)}(x_1, x_2, \ldots, x_n) = F_{X(t_1+T) \ldots X(t_n+T)}(x_1, x_2, \ldots, x_n)$$

$\forall n \in N.$

Exercise. Write down a similar definition for discrete-time random process

Example. Consider i.i.d. discrete-time random process $X_1, X_2, \ldots$. Is this stationary?

Yes!

$(X_{n_1}, X_{n_2}, \ldots, X_{n_k})$ has same distribution as

$(X_{n_1+\tau}, \ldots, X_{n_k+\tau})$.

— If a process is stationary the analysis is usually simpler as the probabilistic properties do not change by time.

However it turns out that not many real-life processes are stationary. Even if a process is stationary it might be difficult to prove it. Fortunately often a weaker notion of stationarity suffices.

Wide-Sense stationary (WSS)

A random process $X(t)$ is called WSS if for all $t, t_1, t_2 \in R$, $T \in R$,

$$\mu_X(t) = \mu_X(t+T)$$

$$R_X(t_1, t_2) = R_X(t_1+T, t_2+T)$$

so $\mu_X(t)$ is a constant and $R_X(t_1, t_2)$ is a function of $t_2 - t_1$.

Example. $X(t) = A \sin(\omega_0 t + \Theta)$ A, Θ are independent & $\Theta \sim$ uniform $[-\pi, \pi]$.

$X(t)$ is WSS because

$$\mu_X(t) = 0 \quad (\text{a constant})$$

$$R_X(t_1, t_2) = \tfrac{1}{2} E[A^2] \cos(\omega_0 (t_1 - t_2))$$

$$= R_X(t_1 - t_2).$$

- For WSS random process it suffices to denote autocorrelation by $R_X(\tau)$.

Exercise. Prove that a stationary process is WSS.

Properties of $R_x(\tau)$

- $R_x(0) = E[x(t)\,x(t)] = E[x^2(t)] \geq 0$.

  $E[x^2(t)]$ is called expected (or average) power in $x(t)$ at time t. For WSS this is not a function of time.

- $R_x(-\tau) = E[x(t+\tau)\,x(t)]$

  $= E[x(t)\,x(t+\tau)]$

  $= R_x(\tau)$.

- $|R_x(\tau)| \leq R_x(0)$ for all $\tau \in R$.

  This follows from Cauchy-Schwarz's inequality,

  $$|E[xy]| \leq \sqrt{E[x^2]\,E[y^2]}$$

  Consider $E[(x-\alpha y)^2] \geq 0$

  $\Rightarrow E[x^2 + \alpha^2 y^2 - 2\alpha xy] \geq 0$

  $\Rightarrow \alpha^2 E[y^2] - 2\alpha E[xy] + E[x^2] \geq 0$

  So discriminant ≤ 0

  $\Rightarrow 4E^2[xy] \leq 4E[x^2]E[y^2]$

$$\Rightarrow |E[XY]| \leq \sqrt{E[X^2]E[Y^2]}.$$

$$|E[x(t)x(t+\tau)]| \leq \sqrt{E[x^2(t)]\,E[x^2(t+\tau)]}$$

$$\Rightarrow |R_x(\tau)| \leq \sqrt{R_x(0)\cdot R_x(0)} = R_x(0).$$

Power Spectral Density (PSD)

Power spectral density is the Fourier transform of $R_x(\tau)$, mathematically

$$S_x(f) = \int_{-\infty}^{\infty} R_x(\tau)\, e^{-j2\pi f\tau}\, d\tau, \quad j = \sqrt{-1}$$

is PSD of $x(t)$.

It can be shown that the inverse Fourier transform of $S(f)$ is $R_x(\tau)$.

$$\int_{-\infty}^{\infty} S_x(f)\, e^{j2\pi f\tau}\, df = R_x(\tau).$$

Properties of PSD

- $S_x(-f) = S_x(f)$.

  This is because $R_x(\tau)$ is even.

- $E[x(t)^2] = R_x(0) = \int_{-\infty}^{\infty} S_x(f)\,df$.

  ↑ Inverse Fourier Transform

  so we get expected or average power in $x(t)$ by integrating the PSD of $x(t)$. This is why $S_x(f)$ is called power spectral density.

- $S_x(f) \geq 0$.

  (Proof omitted)

Lecture 26

(20 November 2023)

Q) (a) Is an event A independent of itself? No in general.

$P(A \cap A) = P(A)^2 \Rightarrow P(A) = 0$ or 1.

(b) Find the probability that exactly one of the events A or B occurs in terms of $P(A)$ $P(B)$ $P(A \cap B)$.

$$P\left((A \cap B^c) \cup (A^c \cap B)\right) = P(A \cap B^c) + P(A^c \cap B)$$

$$= P(A) + P(B) - 2P(A \cap B)$$

(c) Is it true that $P(A \mid A \cup B) \geq P(A \mid B)$?

$$P(A \mid A \cup B) = P(A \mid (A \cup B) \cap B)\, P(B \mid A \cup B) + P(A \mid (A \cup B) \cap B^c)\, P(B^c \mid A \cup B)$$

$$= P(A \mid B)\, P(B \mid A \cup B) + \underbrace{P(A \mid A \cap B^c)}_{=1} P(B^c \mid A \cup B)$$

$$= \lambda P(A|B) + (1-\lambda) 1$$

$$P(A|A\cup B)$$

$$P(A|B) \qquad 1$$

$$\therefore P(A|A\cup B) \geq P(A|B).$$

Q) If X is a positive integer valued RV that satisfies memorylessness property

$$P(X > m+n \mid X > m) = P(X > n), \text{ for any } m, n \in \mathbb{N}.$$

Then prove that X is a geometric random variable.

Proof. Let $P(X > n) = a_n$ for $n \in \mathbb{N}$.

The memorylessness property gives

$$\frac{a_{m+n}}{a_m} = a_n$$

$$\Rightarrow a_{m+n} = a_m a_n \quad \forall\, m, n \in \mathbb{N}$$

$$\Rightarrow a_{m+1} = a_m a_1 = a_1^{m+1}$$

where $a_1 = P(X > 1) = 1 - P(X = 1) \triangleq 1 - p.$

$$P(X = n) = a_{n-1} - a_n = (1-p)^{n-1} - (1-p)^n = (1-p)^{n-1} p.$$

Q) Let X be a discrete RV, Y be a continuous RV and I is a binary RV s.t. X is independent of I

Y is independent of I.

Define

$$Z = \begin{cases} X & \text{if } I=1 \\ Y & \text{if } I=0 \end{cases}.$$

$$Z = \begin{cases} 1 & \text{if } I=1 \\ \text{Uniform}[0,2] & \text{if } I=0 \end{cases}.$$

Z is neither a discrete nor a continuous RV.

$$F_Z(z) = P(Z \le z)$$

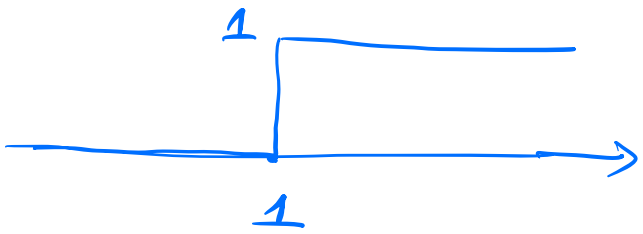
$$= P(Z \le z \mid I=1)P(I=1) + P(Z \le z \mid I=0)P(I=0)$$

$$= p F_X(z) + (1-p) F_Y(z)$$

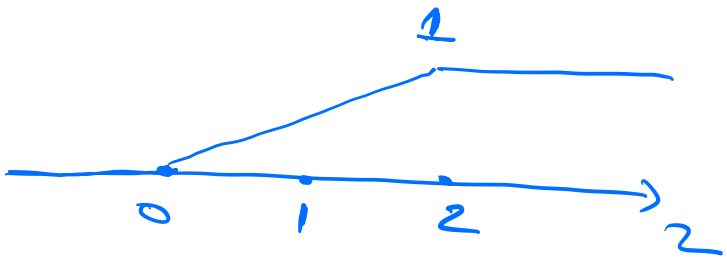
$(p = P(I=1))$

Z is a mixed random variable.

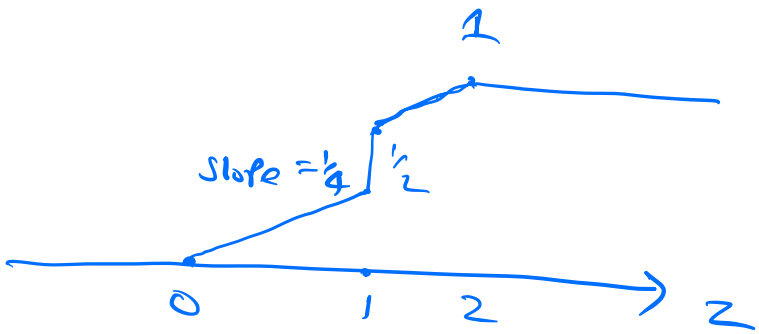
$F_x(z)$



$F_y(z)$



For $p = 1/2$ $f_z(z) =$



Application of Probability in Information Leakage.

The "information content" of a message depends on the degree to which the content of the message is surprising. If a highly likely event occurs, the message carries very little information. If a highly

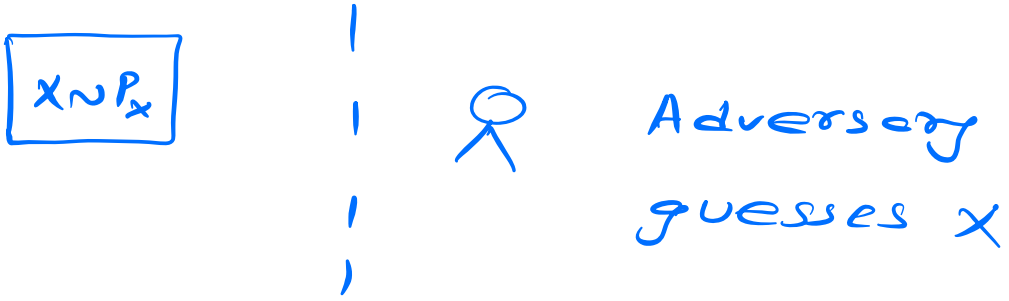
unlikely event occurs the message is more informative.

The information content in an event with probability $p = \log 1/p$.

[all logarithms are to the base 2]

Vulnerability: Suppose $x \sim P_x$. The vulnerability of x is given by

$$V(x) = \max_{x \in \mathcal{X}} P_x(x).$$



$V(x)$ is the worst-case probability that an adversary could guess the value of x correctly.

$\log \frac{1}{V(x)}$ can be viewed as information measure

min-entropy of x $H_\infty(x) = \log \frac{1}{V(x)}$

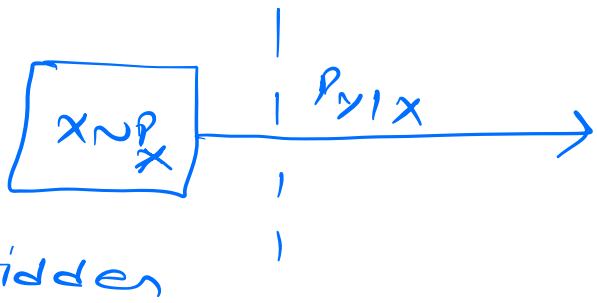
$$= \log \frac{1}{\max_x P_x(x)}.$$

— Given two RVs $(X, Y) \sim P_{XY}$, the conditional vulnerability

$$V(X|Y) = \sum_{y \in \mathcal{Y}} P_Y(y)\, V(X|Y=y)$$

$$= \sum_{y \in \mathcal{Y}} P_Y(y) \max_x P_{X|Y}(x|y).$$

$$H_\infty(X|Y) = \log \frac{1}{V(X|Y)}.$$



information leaked about X?

Information leaked about X from Y

= initial uncertainity in X

—

remaining uncertainity in X after observing Y

Leakage $\mathcal{L}_{P_X}(X \to Y) = H_\infty(X) - H_\infty(X|Y)$

$$= \log \frac{V(X|Y)}{V(X)}.$$

If P_X is unknown the worst-case leakage is of interest.

Theorem. $\max_{P_X} \mathcal{L}_{P_X}(X \to Y) = \mathcal{L}_{P_U}(X \to Y)$

$$= \sum_y \max_x P_{Y|X}(y|x).$$

(Sum of maximum of the columns of $P_{Y|X}$ matrix)

Proof. $2^{\mathcal{L}_{P_X}(X \to Y)} = \frac{V(X|Y)}{V(X)}$

$$= \sum_y P_Y(y) \max_x P_{X|Y}(x|y) \Big/ \max_{x'} P_X(x')$$

$$= \sum_y P_Y(y) \max_x \frac{P_{Y|X}(y|x) P_X(x)}{P_Y(y)} \Big/ \max_{x'} P_X(x')$$

$$= \sum_y \frac{\max_x P_{Y|X}(y|x) P_X(x)}{\max_{x'} P_X(x')}$$

$$\leq \sum_y \max_x P_{Y|X}(y|x) \cdot \frac{\max_{x'} P_X(x')}{\max_{x'} P_X(x')}$$

$$= 2^{L_{P_U}(X \to Y)},$$

where $P_U(x) = \frac{1}{|\mathcal{X}|}, \forall x \in \mathcal{X}$.

Remark. This is an operationally motivated leakage measure and satisfies all the axioms of a leakage measure. Interestingly mutual information $I(X;Y) = H(X) - H(X|Y)$ does not satisfy all.

Q) Let (X,Y) have the joint PDF

$$f_{XY}(x,y) = \begin{cases} x+y, & (x,y) \in [0,1]^2 \\ 0, & \text{otherwise.} \end{cases}$$

Let $Z = X(1+Y)$ $W = X^2$.

$g_1(x,y) = x(1+y)$, $g_2(x,y) = x^2$.

$x = h_1(z,w) = \sqrt{w}$, $y = h_2(z,w) = \frac{z}{\sqrt{w}} - 1$.

$\frac{\partial g_1}{\partial x} = 1+y$, $\frac{\partial g_1}{\partial y} = x$

$\frac{\partial g_2}{\partial x} = 2x$, $\frac{\partial g_2}{\partial y} = 0$.

$$J = \begin{vmatrix} 1+y & x \\ 2x & 0 \end{vmatrix} = 2x^2 = 2w$$

$$f_{zw}(z,w) = \frac{f_{xy}\left(\sqrt{w}, \frac{z}{\sqrt{w}} - 1\right)}{2w}$$

$$= \frac{\sqrt{w} + \frac{z}{\sqrt{w}} - 1}{2w} ,$$

for (z,w) s.t. $0 \le w \le 1$, $\sqrt{w} \le z \le 2\sqrt{w}$.

0 , otherwise.