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, 2^{nd} 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)

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?

<u>Ans.</u> 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.

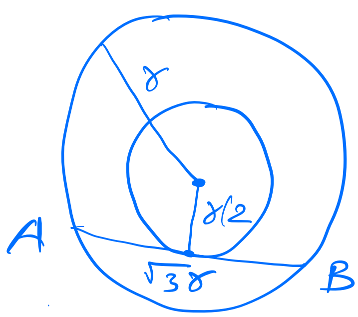
1) It cannot deal with outcomes that are not equally likely; and
2) it cannot handle uncountably infinite outcomes without ambiguity.

(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 $l > \sqrt{3}r$.

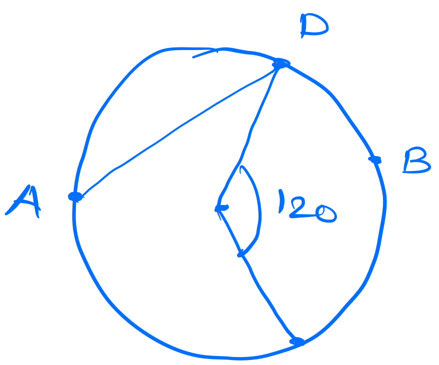
Favourable outcomes for chord center
= all Points inside the circle C_1

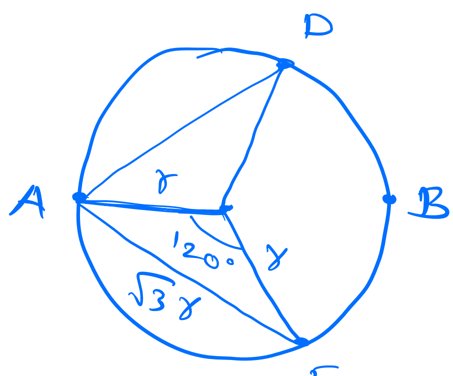
Total possible outcomes for chord center
= all Points inside C

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

(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 $n_E \le n$, we must have $0 \le P(E) \le 1$.

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

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

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)

Probilistic 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]

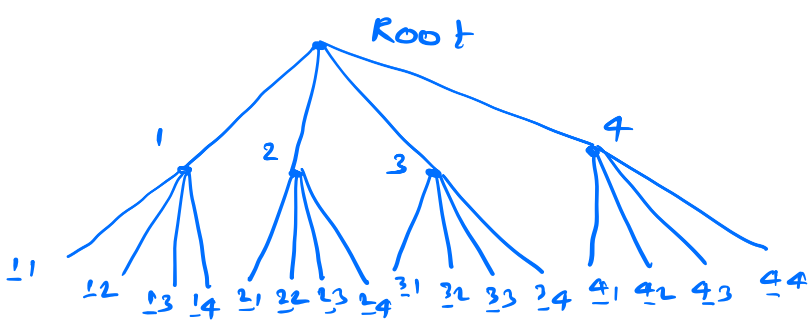
- The elements of the sample space should be
    - mutually exclusive
    - collectively exhaustive

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

[No it is not collectively exhaustive]

<u>Example</u>. Consider rolling a 4-sided die twice (a single random experiment).

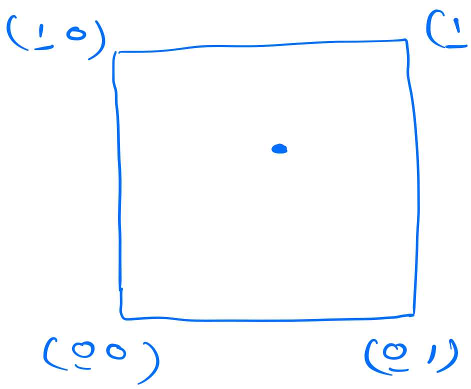
11	12	13	14
21	22	23	24
31	32	33	34
41	42	43	44



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.

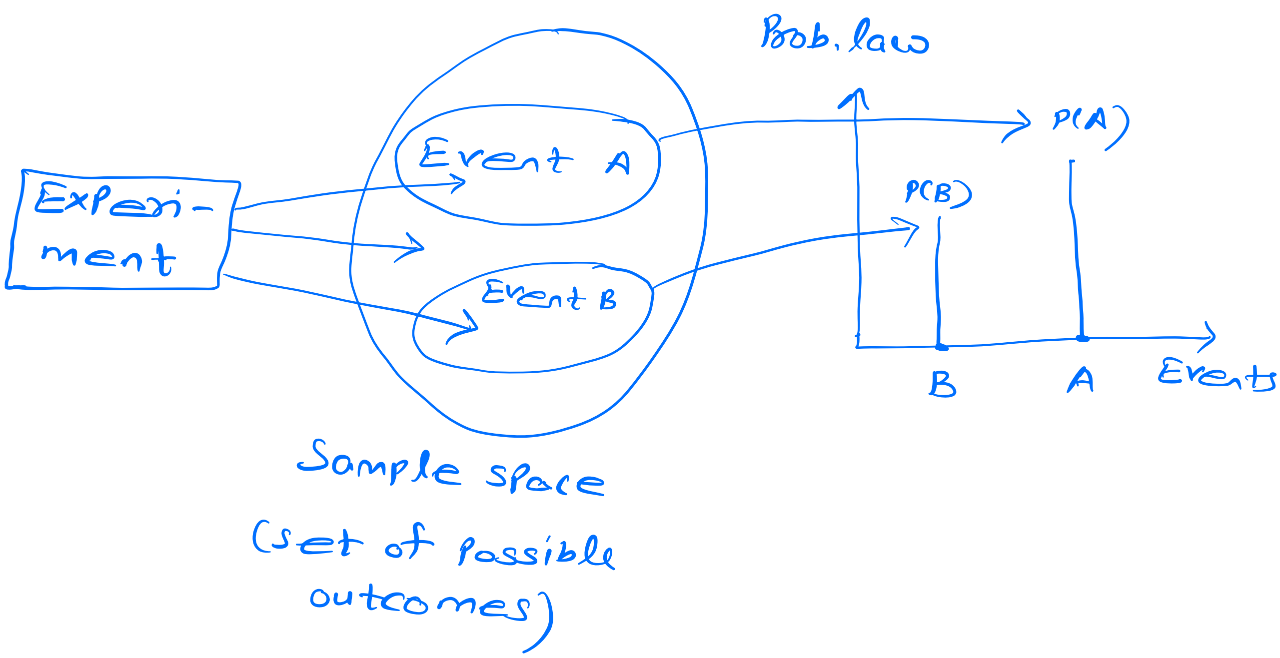
Example. Continuous sample space. Consider throwing a dart on a 1×1 square target.



Record (x,y)

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

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



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.

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

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

1 2 4   1 2 4   1 1 2   1 2 2   1 1 1   1 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
   $$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$ note the infinite number of sets
3. (Normalization) $P(\Omega) = 1$.

Properties of Probability Laws

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

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

(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 $P(A)\le P(B)$.

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

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

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

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

[This proves (1)]

$P(A)=1-P(A^c)\le 1$ since $P(A^c)\ge 0$ by nonnegativity

[This proves (2)]

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

$\implies P(\phi)=0$. [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)]

$$\boxed{A \subseteq B \qquad B = A \cup (B\setminus A) = A \cup (B\cap A^c)}$$

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

$$P(B) = P(A\cup(B\cap A^c)) = P(A) + P(B\cap A^c) \geq 0.$$

[This proves (5)]

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

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

just write everything as union of disjoint sets

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

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

[This proves (6)]

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

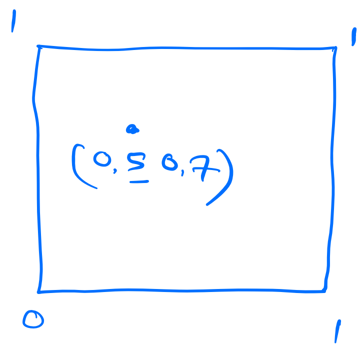
$$\leq P(A) + P(B) \quad (\because \text{Nonnegativity})$$

[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 1×1 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.

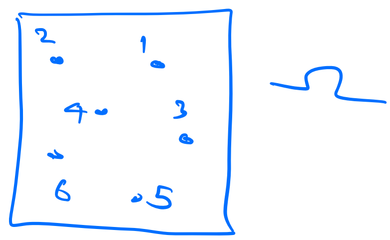
- More precisely given an experiment, a corresponding sample space, and a probability law, suppose that we know that the outcome is within some given event B, we wish to quantify the likelihood that the outcome ~~also~~ belongs to some other given event A.

- 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.

<u>Exercise 3.1</u>. 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

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

<u>Exercise 3.3</u> 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:

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

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

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

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

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

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

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

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

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

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

Bayes' Rule (Refined version)

Let $A_1, A_2, \ldots, A_n$ 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

$$P(A_i \mid B) = \frac{P(A_i) P(B \mid A_i)}{P(B)}$$

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

Example. (Exercise 3.4 in class)

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

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

we are given that

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

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

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

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

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

$P(A|B) = P(A)$ 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:

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

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

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

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

(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;

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

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

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

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

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

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

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

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

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

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

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 independence and vice versa.

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

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

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$\}$.

$$\boxed{\begin{array}{l} P(H_1 \cap H_2 | F) = P(H_1|F)P(H_2|F) \quad \text{and} \\ P(H_1 \cap H_2) \neq P(H_1)P(H_2). \end{array}}$$

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.

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

Partitions: consider n and $n_1, \dots, n_r$ s.t.

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

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

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

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.

$$B_1 = A_1$$

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

(i) B_i's are disjoint (verify)

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

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

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

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

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

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

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

$\blacksquare$

Consequences.

1) Let $A_1 A_2 ---$ 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 consequency by considering complements of the events and using DeMorgan's laws.

Lecture 5

(19 August 2023)

Continuity of probability

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

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

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 $A_i \subseteq \bigcup_{j=1}^{i'-1} A_j$

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

so things might not be possible to write 100% in purely math terms, sometimes you have to use language at a lot of places in these probabilty proofs >>

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

base case is for n = 2, n=1 is trivial.

$C_2 = A_1 \cup A_2 = B_1 \cup B_2$ (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 $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$.

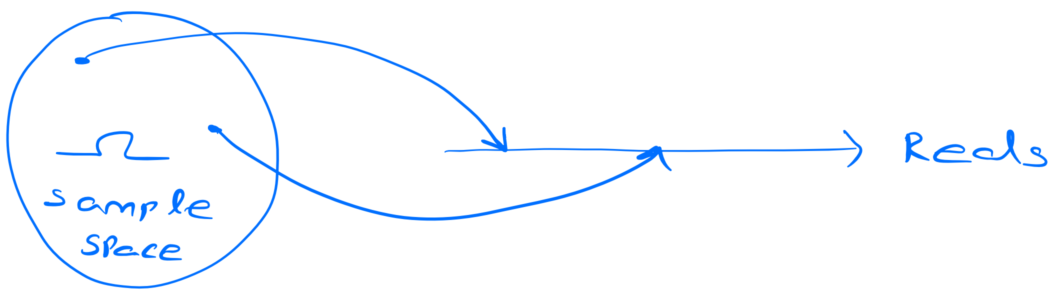
Let $x \in \bigcup_{i=1}^{\infty} A_i \Rightarrow \exists\, n \in \mathbb{N}$ s.t. $x \in A_n$

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

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 $\quad \Omega = \{H, T\}$

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

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

Claim. $\sum_{x \in \mathcal{X}} P_X(x) = 1.$

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

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

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

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

(by additivity) (normalization)

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

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

Example. 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 $1-p$.

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

$$p_X(1) = p \qquad p_X(0) = 1-p.$$

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.

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

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

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.

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

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

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

Proof:

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

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

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

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

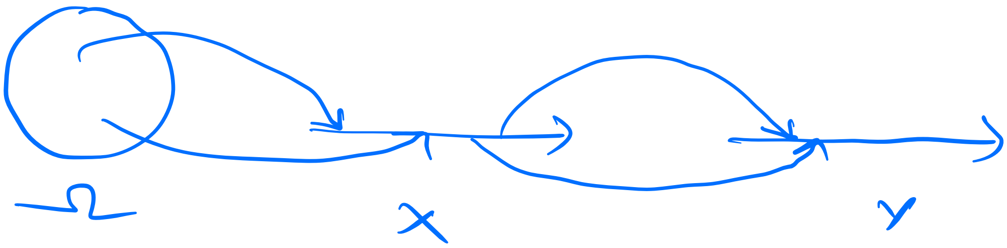
$$= \binom{100}{5}(0{,}01)^5(0{,}99)^{95} = 0{,}00290.$$

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., $X : \Omega \to \mathbb{R}$, and Y be a function of X, i.e., $Y = g(X)$, $g: \mathbb{R} \to \mathbb{R}$.



Y is also a random variable.

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

$\rightarrow$ $\boxed{\text{Exercise 6.1 Prove this}}$

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

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

Find the PMF of Y.

Proof. $\mathcal{X} = [-4 : 4]$,

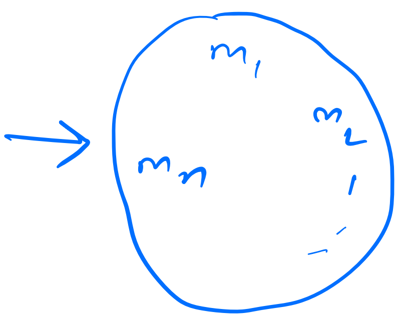
$\mathcal{Y} = [0 : 4]$.

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

$p_Y(y) = 0$, o.w.

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. $k = \sum_{i=1}^{n} k_i$.

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 $k \to \infty$,

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

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

$$E[X] = 1.p + 0.(1-p) = p.$$

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)

<u>Variance</u>. $Var(X) = E[(X - E[X])^2]$.

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.

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

$E[X] = 0$.

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

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

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

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

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

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

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

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

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

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

Using this

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

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

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

Properties of mean and variance.

- $Y = aX + b$

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

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

$$= E\left[\left(aX \cancel{+b} - aE[X] \cancel{-b}\right)^2\right]$$

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

Variance in Terms of Moments

$$Var(X) = E[X^2] - (E[X])^2.$$

Mean and Variance of some common RVs

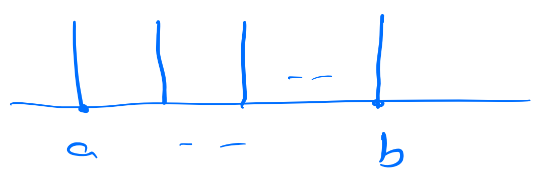
Bernoulli RV:

$$p_X(1) = p = 1 - p_X(0),$$

$E[X] = p$, $E[X^2] = p$, $Var(X) = p - p^2 = p(1-p)$,

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

since $\operatorname{Var}(X) = \operatorname{Var}(X-a+1) = \operatorname{Var}(Y)$,

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

Binomial random variable:

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

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

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

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

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

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

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

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

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

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

$$[\, n' = n-1 \quad k' = k-1 \,]$$

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

$$= np\,(1 + (n-1)p)$$

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

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

$$= np(1 + \cancel{np} - p - \cancel{np}) = np(1-p).$$

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

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

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

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}.

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

$$p_{XY}(x,y) = P(\{X=x\} \cap \{Y=y\})$$
$$=: P(X=x, Y=y).$$

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

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

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

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

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

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

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

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

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

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

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

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,

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

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

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

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

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

Example. In each lecture a professor is asked 0, 1 or 2 questions with equal probability 1/3. He answers each question incorrectly with probability 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 $P_{XY}(x,y)$.

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

$P_{Y|X}(0|1) = 3/4$, $P_{Y|X}(1|1) = 1/4$, $P_{Y|X}(2|1) = 0$,

$P_{Y|X}(0|2) = 9/16$, $P_{Y|X}(1|2) = 3/8$, $P_{Y|X}(2|2) = 1/16$.

P.T.O

Conditional Expectation

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

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

expectation of R conditioned on an event A

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

expectation of a function of a RV conditioned on event A

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

expectation of a RV conditioned on another random variable

Total Expectation Theorem:

If $A_1, \ldots, A_n$ 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. $E[(X-f(Y))^2]$

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

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

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

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

Conditional Variance

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

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

Exercise. Prove that

$$\mathrm{Var}(X) = E[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(E[X \mid Y]).$$

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.

$x \backslash y$	1	2	3	4
1	0	$1/20$	0	0
2	0	$1/20$	$3/20$	$1/20$
3	$2/20$	$4/20$	$1/20$	$2/20$
4	$1/20$	$2/20$	$2/20$	0

p_{XY}

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 $A = \{X \geq 3,\ Y \leq 2\}$?

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

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

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

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

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

Exercise

Two events A and B are independent if and only if $\mathbb{1}_A$ and $\mathbb{1}_B$ are independent RVs.

Independence of several Random Variables

Three RVs X, Y and Z are said to be independent if

$$P_{XYZ}(x,y,z) = P_X(x)\, P_Y(y)\, P_Z(z) \text{ for all } x, y, z.$$

Independence of n RVs:

$$P_{X_1 X_2 \ldots X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} P_{X_i}(x_i).$$

Variance of sum of Independent RVs

$$Var(X+Y)$$

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

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

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

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

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

where $Cov(X,Y) = E[(X-E[X])(Y-E[Y])]$.

If X and Y are independent, $Cov(X,Y)=0$.

$$Var(X+Y) = Var(X) + Var(Y).$$

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

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

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

Let $X_i = \mathbb{1}\{i\text{th coin toss is heads}\}$, $i \in [1:n]$.

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

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

Lecture 10

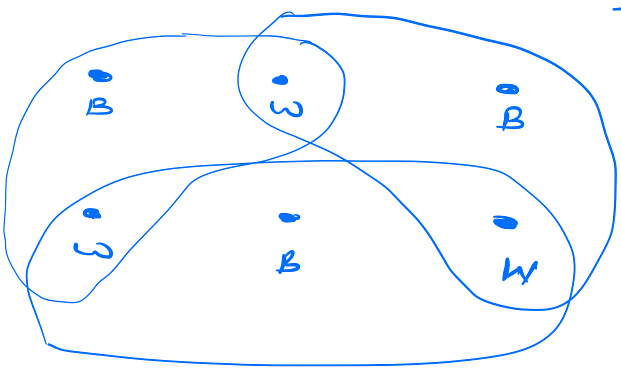
(7 September 2023)

Some Applications

Combinatorics and Graph Theory

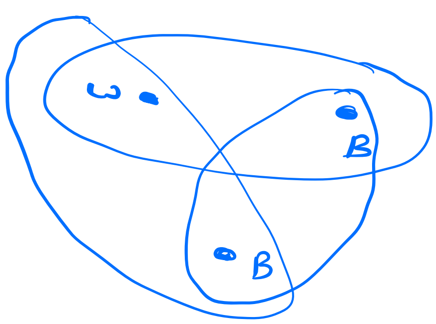
2-Coloring: [Existence proof using union bound]

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. $|T_i| = \ell$ for $i \in [1; m]$, there exists a valid 2-coloring of S such that no T_i is monochromatic if $m < 2^{\ell-1}$.

For the examples above

$$\ell = 3 \quad m = 3 \implies 3 < 2^{3-1} = 7$$

$$\ell = 2 \quad m = 3 \implies 3 \not< 2^2 - 1 = 3$$

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 $\frac{1}{2}$.

Let E_i be the event that T_i is monochromatic.

B W

either black or white

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

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

$P($ each T_i is not monochromatic$)$

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

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

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).

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

we have to show that at least one of them has a length of m/2

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

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

Number the edges from 1 through $\underline{m}$ and let

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

= number of subgraphs which has one end in A and other at B / total number of subgraphs.

$$P(X_i = 1) = \frac{2^{n-2}}{2^{n-1}-1},$$ 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.

this time in split, one of the half can be empty as well.
2^(n-2)/2 * 2 is answer. *2becasue one half goes with one node or the other half goes with that node,there are two nodes.

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

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

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

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

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

- 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

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

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 $H(X_2) \le H(X_1)$? Yes!

$$H(X_2) = h(p) = -p\log p - (1-p)\log(1-p)$$

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

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

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

$h(\frac{1}{2}) = 1$ 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.

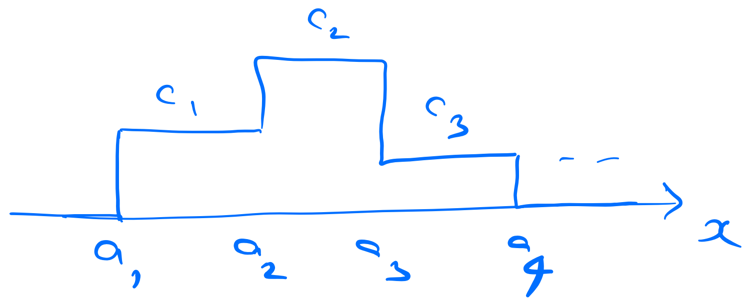
Consider a PDF

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

$$\int_a^b f_X(x)\,dx = 1 \implies c = 1/_{b-a}.$$

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

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.

$Y = g(X)$.

$$Y = g(X) = X \quad \text{continuous RV}$$

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

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

Properties

- $Y = aX + b$

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

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

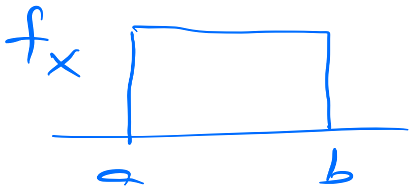
- $\mathrm{Var}(X) = E[(X - E[X])^2]$

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

Example: Mean and variance of the Uniform Random variable.

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

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



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

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

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

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

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

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

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

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

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

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

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

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

$$Var(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = 1/\lambda^2.$$

Lecture 12

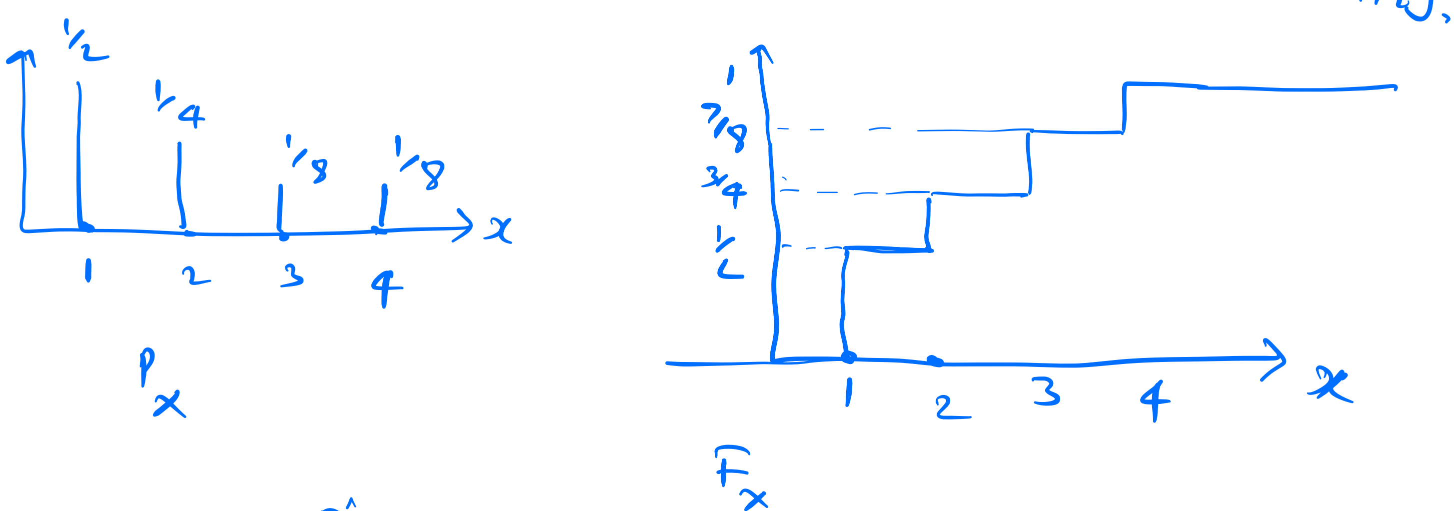
(14 September 2023)

Cumulative Distribution Functions

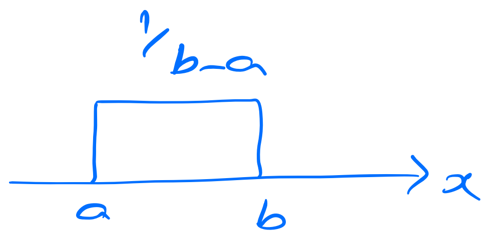
We would like to describe all kinds of RVs with a single mathematical concept. This is accomplished with CDF.

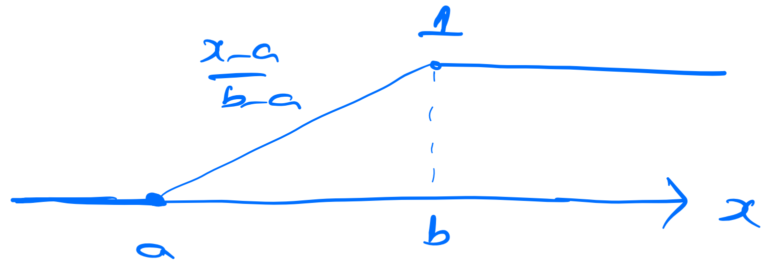
$$F_X(x) = \begin{cases} \sum_{k \le x} p_X(k) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{x} f_X(t)\,dt & \text{if } X \text{ is continuous.} \end{cases}$$

- 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 therefore has a well-defined probability.



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 $x \le y$ then $F_X(x) \le F_X(y)$.

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

$$\le P(X \le y) = F_Y(y).$$

2) $\lim\limits_{x \to -\infty} F_X(x) = 0$ $\qquad \lim\limits_{x \to \infty} F_X(x) = 1.$

$$A_n = \{X \le n\} \quad A_1 \subseteq A_2 \subseteq \dots$$

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

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

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

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

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

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

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

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

3) F_X is right continuous.

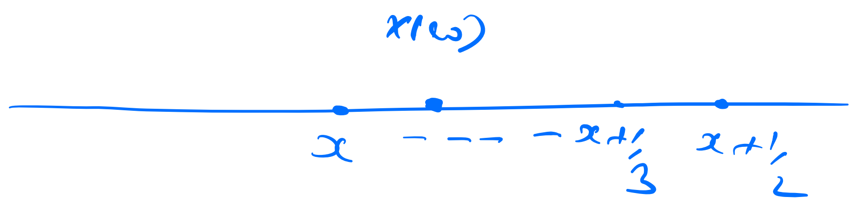
$$\lim_{\varepsilon\to 0+} F_X(x+\varepsilon) = F_X(x).$$

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

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

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

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



$$X(\omega) \leq x + 1/i, \ \forall i \in [1:\infty] \implies X(\omega) \leq x,$$

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

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

- 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,

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

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

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

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

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

$$= F_X(x).$$

4) If X is discrete and takes integer values,

$$F_X(k) = \sum_{i=-\infty}^{k} P_X(i),$$

$$P_X(k) = P(X \leq k) - P(X \leq k-1)$$

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

If X is continuous,

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

$$f_X(x) = \frac{d F_X(x)}{dx}, \quad \text{for all } x \text{ s.t. } f \text{ is continuous at } x.$$

[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 $P_{X_i}(k) = 1/10,\ k \in [1:10]$.

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.

actually it is a two way proof.

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

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

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

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

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

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

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

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

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

<u>Solution</u>,

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

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

Xi is either 0 or 1, so its square is also Xi only.

$$= \sum_{i=1}^{n} E[X_i] + \sum_{i \neq j} E[X_k X_j]$$

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

remember X is sum of all Xis

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

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

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

$$= \sum_{i=1}^{n} P_{X_i}(1)\, E[X \mid X_i = 1].$$

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 $1/2$ and 1 with probability $1/2$.

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

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

(d) Find $Var(Y)$.

Solution. $X_1, X_2, \ldots, X_n$ 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

$Y_i = X_j + X_k$ 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

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

$$\therefore E[Y_i Y_j] = E[Y_i] E[Y_j] = \frac{1}{4}.$$

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

$$= \sum_{i=1}^{m} E[Y_i^2] + \sum_{i \neq j} E[Y_i Y_j] - \left(m \frac{1}{2}\right)^2$$

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

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

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. $f_X(x) = \begin{cases} 0.5 & \text{for } 0 \le x \le 0.5 \\ 0 & \text{for } 0.5 < x \le 1. \end{cases}$

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.

$$P(X \leq 1) = 0.6 \geq \tfrac{1}{2}$$

$$P(X \geq 1) = 0.8 \geq \tfrac{1}{2}$$

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

$$P(X \leq 2) = 1 \geq \tfrac{1}{2}$$

$$P(X \geq 2) = 0.6 \geq \tfrac{1}{2}$$

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

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

$$\geq \tfrac{1}{2}$$

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

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

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

$$P(X \leq \alpha) = 0.5 \geq \tfrac{1}{2}$$

$$P(X \geq \alpha) = \int_1^\infty f_X(x)\,dx = \tfrac{1}{2} \geq \tfrac{1}{2}$$

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 $Var(x^2 y)$.

$$Var(x^2 y) = E[x^4 y^2] - (E[x^2 y])^2$$

$$= 2 \cdot 1 - (1 \cdot 0)^2 = 2.$$

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

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

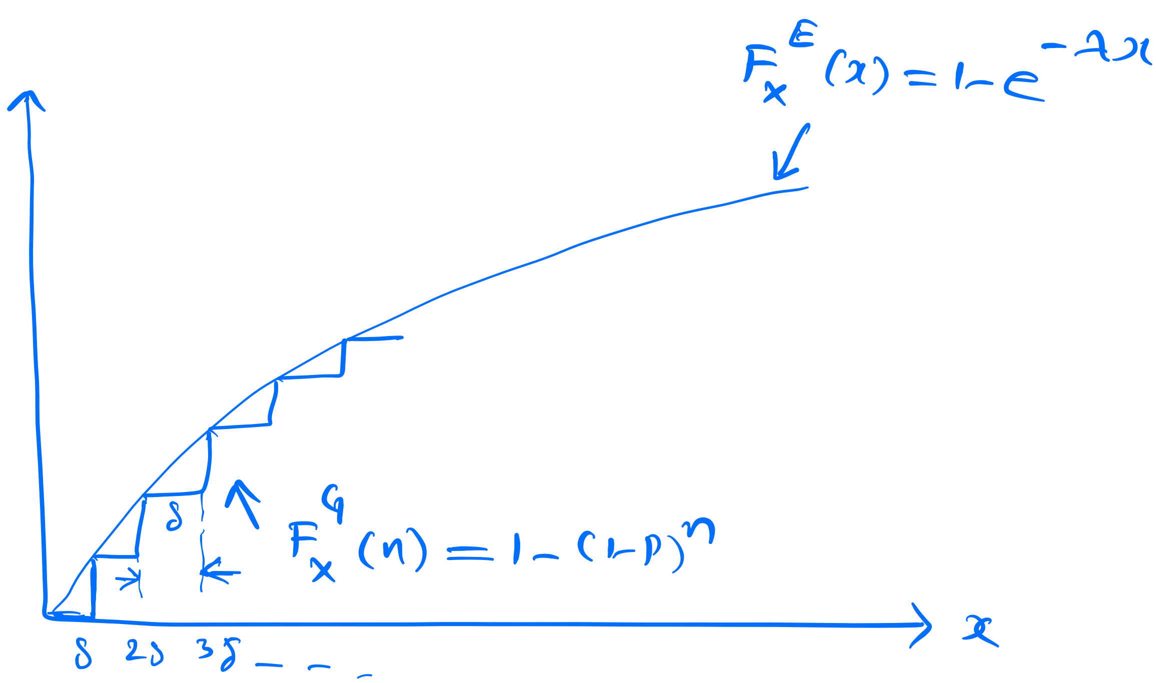
Consider the values of $F_X^E(x)$ at $x = n\delta$ $n = 1, 2, \ldots$

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

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

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.,

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



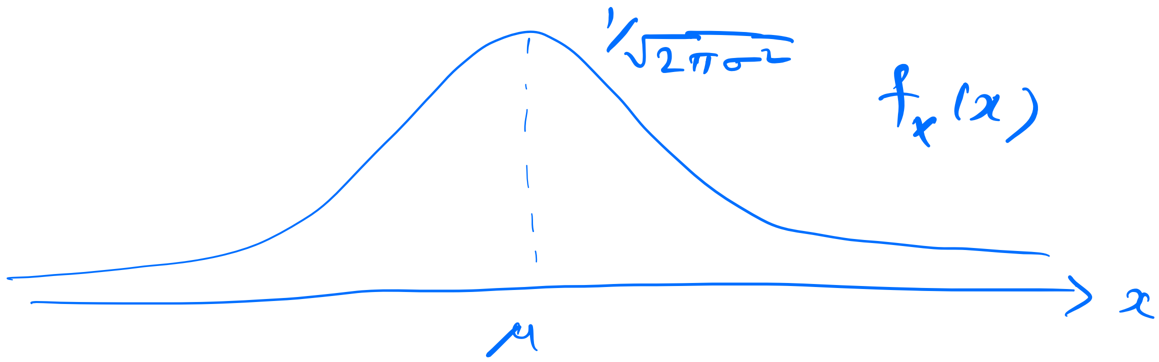
Suppose now we toss a biased coin very quickly (every δ seconds where $\delta \ll 1$) with a small probability of heads (equal to $p = 1 - e^{-\lambda\delta}$). Then the first time to obtain a head (a geometric random variable with parameter p) is a close approximation to an exponential r.v. with parameter 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 $t = \dfrac{x-\mu}{\sigma}$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $y = ax + b$.

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

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

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

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

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

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

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

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

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

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

$$P(X \ge 80) = P\left(\frac{X - 60}{20} \ge \frac{80 - 60}{20} = 1\right)$$

$$= P(Y \ge 1) = \Phi(1).$$

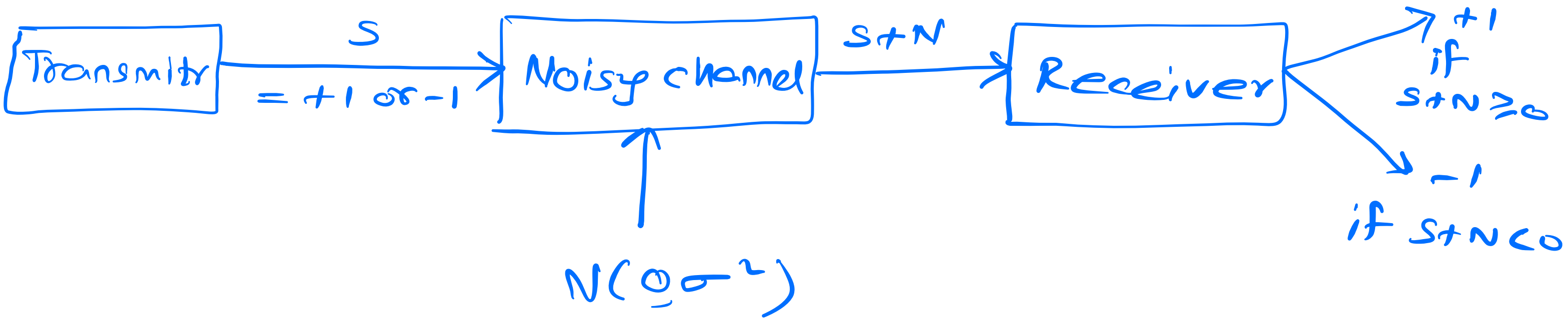
In general

$$P(X \le x) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

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 than -1 so that $s+N = 1+N < 0$.

When $s=-1$ $P(N \geq 1)$ is the error probability.

$$P(N \geq 1) = 1 - P(N<1) = 1 - P\left(\frac{N}{\sigma} < \frac{1}{\sigma}\right)$$

$$= 1 - \Phi(1/\sigma).$$

When $s=+1$ $P(N \leq -1) = P(N \geq 1)$

$$= 1 - \Phi(1/\sigma).$$

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 $\mathbb{R}^2$.

<u>Exercise</u>. 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 R^2.

In particular when B is a rectangle

$$P(a \le X \le b,\ c \le Y \le d) = \int_{y=c}^{d} \int_{x=a}^{b} f_{XY}(x,y)\,dx\,dy.$$

$$B = R^2 \Rightarrow \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_X(x,y)\,dx\,dy = 1.$$

To interpret the joint PDF we let δ be a small positive number and consider the probability of a small rectangle.

$P(a \le X \le a+\delta,\ c \le Y \le c+\delta)$

$$= \int_c^{c+\delta} \int_a^{a+\delta} f_{XY}(x,y)\,dx\,dy \approx f_{XY}(a,c)\,\delta^2$$

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 $P(X \in A) = \int_A f_X(x)\,dx$

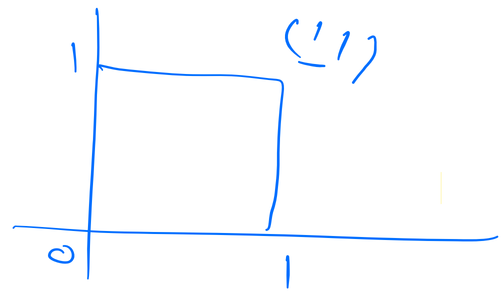
marginal PDF f_X of X is given by

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

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

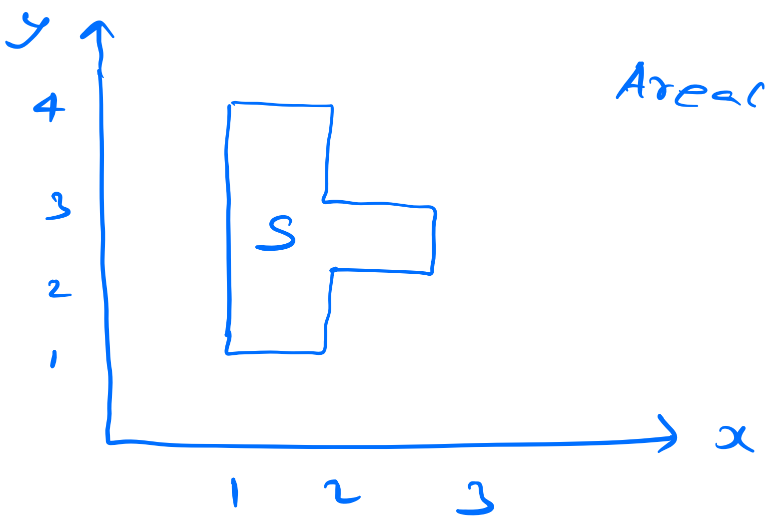
Example (Two-Dimensional Uniform PDF).

$$f_{XY}(x,y) = \begin{cases} c & \text{if } 0 \le x \le 1 \text{ and } 0 \le y \le 1 \\ 0 & \text{o.w} \end{cases}$$



where c is a constant.

Example.



Area(S) = 4

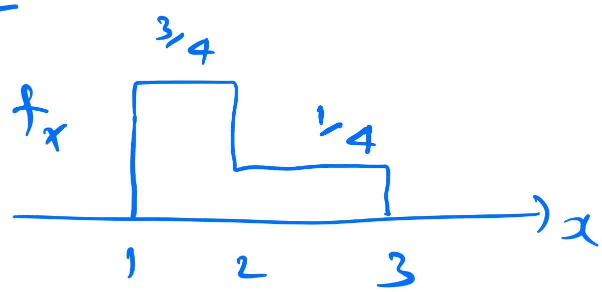
$$f_{XY}(x,y) = \begin{cases} 1/4 , & \text{if } (x,y) \in S \\ 0 & \text{o.w.} \end{cases}$$

Find f_X , f_Y.

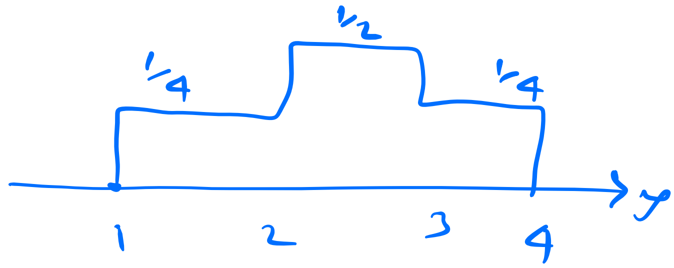
$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy$$

$$= \begin{cases} \displaystyle\int_1^4 \tfrac{1}{4}\, dy & x \in [1,2] \\ \displaystyle\int_2^3 \tfrac{1}{4}\, dy & x \in [2,3] \end{cases}$$

$=$



$f_Y =$



Joint CDFs. If X and Y are two RVs associated with the same experiment we define their joint CDF by

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

For continuous RVs X and Y

$$F_{XY}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(s,t)\, dt\, ds.$$

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

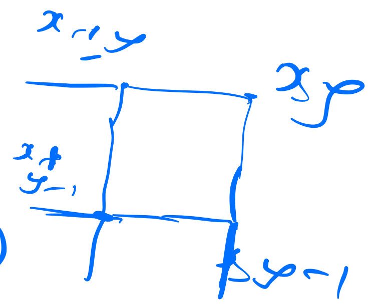
$$P(x_1 < X \le x_2,\ y_1 < Y \le y_2)$$

$$= F_{XY}(x_2, y_2) - F_{XY}(x_1, y_2) - F_{XY}(x_2, y_1) + F_{XY}(x_1, y_1)$$

For discrete RVs X and Y

$$F_{XY}(x, y) = \sum_{l \le x} \sum_{k \le y} P_{XY}(l, k).$$

$$P_{XY}(x, y) = F_{XY}(x, y) - F_{XY}(x-1, y) - F_{XY}(x, y-1) + F_{XY}(x-1, y-1)$$



<u>Example</u>. 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 $B \subseteq \mathbb{R}^3$.

$\Longrightarrow$ 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 $\underline{x}$ 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 $A \subseteq \Omega$.

$$B = \mathbb{R} \implies \int_{-\infty}^{\infty} f_{X|A}(x)\, dx = 1.$$

Suppose A is of the form $A = \{X \in C\}$.

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

$$= \frac{1}{P(X \in C)} \int_{B \cap C} f_X(x)\, dx$$

$$= \int_B \frac{f_X(x)\, \mathbb{1}\{x \in C\}}{P(X \in C)}\, dx$$

so $f_{X|X\in C}(x) = \begin{cases} \dfrac{f_X(x)}{P(X\in C)}, & \text{if } x\in C, \\ 0, & \text{o.w.} \end{cases}$

Conditional CDF

$$F_{X|A}(x) = P(X\le x \mid A).$$

Theorem. Let $A_1, A_2, \dots, A_n$ 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

$$P(x \le X \le x+\delta_1 \mid y \le Y \le y+\delta_2)$$

$$= P(x \le X \le x+\delta_1,\ y \le Y \le y+\delta_2) \,/\, P(y \le Y \le y+\delta_2)$$

$$\approx \frac{f_{XY}(x,y)\,\delta_1\cancel{\delta_2}}{f_Y(y)\,\cancel{\delta_2}}.$$

In view of the above,

Define $P(X \in A \mid Y = y) = \int_A f_{X|Y}(x|y)\, dx$.

Interpretation:

$$P(X \in A \mid Y = y) \stackrel{\cdot}{=} \lim_{\delta \to 0} P(X \in A \mid y \le Y \le y + \delta)$$

$$= \lim_{\delta \to 0} \frac{P(X \in A, y \le Y \le y+\delta)/\delta}{P(y \le Y \le y+\delta)/\delta}$$

$$= \lim_{\delta \to 0} \frac{\int_{x \in A} \int_y^{y+\delta} f_{XY}(x, t)\, dt\, dx}{\delta} \Big/ \int_y^{y+\delta} f_Y(y)\, dy$$

$$= \int_{x \in A} f_X(x) \left[\lim_{\delta \to 0} \int_y^{y+\delta} f_{Y|X}(t|x)\, dt \Big/ \delta \right] dx \Big/ \lim_{\delta \to 0} \int_y^{y+\delta} f_Y(y)\, dy$$

$$= \int_{x \in A} f_X(x) \frac{d}{dy} F_{Y|X}(y|x) \Big/ \frac{d}{dy} F_Y(y)\, dx$$

$$= \int_{x \in A} \frac{f_X(x)\, f_{Y|X}(y|x)}{f_Y(y)}\, dx.$$

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

<u>proof</u>. $\int_{-\infty}^{\infty} E[X|Y=y] f_Y(y)\, dy$

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

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

— The total expectation theorem can often be used to calculate mean and variance.

<u>Example</u>. 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 $U \sim \text{Uniform}[a,b] \Rightarrow E[U] = \frac{a+b}{2}$,

$$E[U^2] = \frac{a^2+ab+b^2}{3}.$$

$E[X|A_1] = 1/2$, $E[X|A_2] = 3/2$

$E[X^2|A_1] = 1/3$, $E[X^2|A_2] = 7/3$.

$E[X] = 7/6$, $E[X^2] = \frac{15}{9}$.

$Var(X) = E[X^2] - E[X]^2 = 11/36$.

Independence

Two continuous RVs are independent if

$$f_{XY}(x,y) = f_X(x) f_Y(y) \quad \forall x, y.$$

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

$$F_{XY}(x,y) = F_X(x) F_Y(y), \quad \text{for all } x, y.$$

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

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

$$\Rightarrow \quad f_{XY}(x,y) = f_X(x) f_Y(y).$$

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

Lecture 17

(12 September 2023)

The Continuous Bayes' Rule

$$X \sim f_X \longrightarrow \boxed{f_{Y|X}} \longrightarrow Y$$

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 $f_{X|Y}(x|y)$.

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

$$\Longrightarrow f_{X|Y}(x|y) = \frac{f_X(x)\, f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(t)\, f_{Y|X}(y|t)\, dt}$$

Inference about a Discrete RV:

$$P(A|Y=y) = \lim_{\delta \to 0} P(A \mid Y \in [y, y+\delta])$$

$$= \lim_{\delta \to 0} \frac{P(A)\, P(Y \in [y, y+\delta] \mid A)/\delta}{P(Y \in [y, y+\delta])/\delta}$$

$$= \frac{P(A)\, f_{Y|A}(y)}{f_Y(y)}.$$

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

$$X \sim P_X \longrightarrow \boxed{f_{Y|X}(y|x)} \longrightarrow Y$$

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

$$= \frac{p_X(x)\, f_{Y|X}(y|x)}{\sum_{x=0}^{M-1} p_X(x)\, f_{Y|X}(y|x)}, \quad \text{where} \quad \sum_{x=0}^{M-1} p_X(x) = 1.$$

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 Bayes' 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

$$P_{X|Y}(\hat{x}_{MAP}(y)|y) \geq P_{X|Y}(\hat{x}_A(y)|y),$$

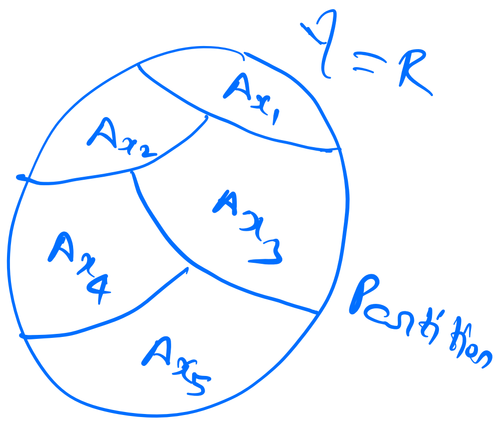
for all A and y.

- 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.



$$P(\hat{x}_A(Y) = X) = \sum_{x=0}^{M-1} P_X(x)\, P(\hat{x}_A(Y) = x \mid X = x)$$

$$= \sum_{x=0}^{M-1} P_X(x)\, P(Y \in A_x \mid X = x)$$

$$= \sum_{x=0}^{M-1} P_X(x) \int_{y \in A_x} f_{Y|X}(y|x)\, dy$$

$$= \int_{y=-\infty}^{\infty} P_X(\hat{x}_A(y))\, f_{Y|X}(y \mid \hat{x}_A(y))\, dy$$

$$= \int_{y=-\infty}^{\infty} P_{X|Y}(\hat{x}_A(y) \mid y)\, f_Y(y)\, dy \longrightarrow ①$$

Similarly

$$P(\hat{x}_{MAP}(Y) = X) = \int_{y=-\infty}^{\infty} p_{X|Y}(\hat{x}_{MAP}(y)|y) f_Y(y)\, dy \rightarrow (2)$$

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), 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,

$$\frac{f_{Y|X}(y|1)\,P_X(1)}{\cancel{f_Y(y)}} \underset{\hat{x}(y)=0}{\overset{\hat{x}(y)=1}{\gtrless}} \frac{f_{Y|X}(y|0)\,P_X(0)}{\cancel{f_Y(y)}},$$

equivalently

$$\Lambda(y) = \frac{f_{Y|X}(y|1)}{f_{Y|X}(y|0)} \underset{\hat{x}(y)=0}{\overset{\hat{x}(y)=1}{\gtrless}} \frac{P_0}{P_1} = \eta.$$

$\uparrow$
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 minds

- 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

$$A_1 = \{y : \Lambda(y) \geq \eta\} \qquad A_0 = \{y : \Lambda(y) < \eta\}$$
$$\downarrow \qquad\qquad\qquad \downarrow$$
$$\hat{x} = 1 \qquad\qquad\qquad \hat{x} = 0.$$

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 $\quad P(\text{Er} \mid X=1) = \int_{y \in A_0} f_{Y|X}(y|1)\, dy.$

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)

$P_X(b) = P_1 \quad P_X(-b) = P_0. \quad Y = X + Z, \quad Z \sim N(0, \sigma^2),$

X and Z are independent.

$$P(Y \in B \mid X = x) = P(X + Z \in B \mid X = x)$$

$$= P(Z + x \in B)$$

$\therefore$ $Y \mid X = b \sim N(b, \sigma^2)$ $\quad$ $Y \mid X = -b \sim N(-b, \sigma^2)$,

$$f_{Y|X}(y \mid b) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-b)^2/2\sigma^2}$$

$$f_{Y|X}(y \mid -b) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y+b)^2/2\sigma^2}.$$

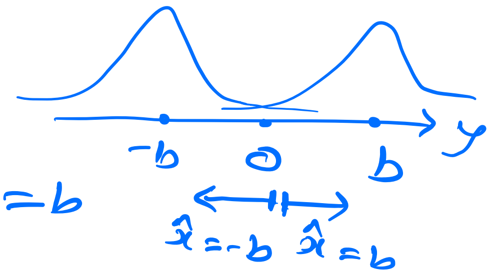
MAP rule:

$$\Lambda(y) = \frac{f_{Y|X}(y \mid b)}{f_{Y|X}(y \mid -b)} = e^{\frac{(y+b)^2 - (y-b)^2}{2\sigma^2}}$$

$$= e^{2yb/\sigma^2} \underset{\hat{x}(y) = -b}{\overset{\hat{x}(y) = b}{\gtrless}} \frac{p_0}{p_1} = \eta.$$

$$\Rightarrow \quad y \underset{\hat{x}(y) = -b}{\overset{\hat{x}(y) = b}{\gtrless}} \frac{\sigma^2}{2b} \log \eta.$$

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 $y \geq 0 \longrightarrow \hat{x} = b$

$y < 0 \longrightarrow \hat{x} = -b.$

Coming back to the case with $b > 0$

$$P(Er \mid X = b) = P(Y < 0 \mid X = b)$$

$$= P\left(\frac{Y-b}{\sigma} < \frac{-b}{\sigma} \;\middle|\; X = b\right)$$

$$= P\left(N < \frac{-b}{\sigma}\right)$$

$$= P\left(N \geq \frac{b}{\sigma}\right).$$

$$P(Er \mid X = -b) = P\left(N > \frac{b}{\sigma}\right)$$

$$= 1 - P(N \leq b/\sigma).$$

$$P(\text{Error}) = P(N \geq b/\sigma).$$