

Lecture 1: Introduction

Satyajit Thakor
IIT Mandi

17 February, 2020

Why study probability?

- ▶ Probability is the logic of uncertainty.
- ▶ Statistics is concerned with collection, organization, analysis, interpretation and presentation of data.

Probability is extremely useful in a wide variety of fields, such as

- ▶ Physics: Quantum physics, statistical mechanics
- ▶ Biology: study of random mutations of genes
- ▶ Computer science: Randomized algorithms
- ▶ Meteorology: Weather forecast
- ▶ Finance and gambling: Modeling stock prices
- ▶ Political science: Analysis of public opinion, prediction
- ▶ Medicine: Randomized clinical trials.

Introduction to IC252

- ▶ IC252 - a foundation course on which you will build understanding for other advanced courses.

Topics:

- ▶ Probability (6 lectures)
- ▶ Random variables (9 lectures)
- ▶ Measures of central tendency, dispersion and association (11 lectures)
- ▶ Statistics (14 lectures)
- ▶ Case study (2 lectures)

Introduction to IC252

Reference books:

- ▶ Sheldon Ross, *Introduction to Probability and Statistics for Engineers*, 5/e (2014), Elsevier
- ▶ Morris H. DeGroot and Mark J. Schervish, *Probability and Statistics* (4/e)(2012), Addison- Wesley.
- ▶ Blitzstein and Hwang, *Introduction to Probability* (2015), CRC Press.
- ▶ William Feller, *An Introduction to Probability*, (3/e) (2008), Volume 1, Wiley.
- ▶ Freedman, Pisani, Purves, *Statistics* (4/e)(2014), W. W. Norton & Company.

Introduction to IC252

Evaluation (Theory 70%, Lab 30 %):

- ▶ Theory (approx. %): Quiz-1 (20%), Quiz-2 (20%), Endsem exam (50%), Assignments (10%)
- ▶ Lab (approx. %): Weekly evaluations (33%), Mid-term exam (27%), End-term exam (40%)

Lecture hours:

- ▶ Monday (10:00 – 10:50 AM), Tuesday (09:00 – 09:50 AM), Friday (08:00 – 08:50 AM)

Lab hours:

- ▶ Tuesday, Thursday and Friday, 2:00 PM – 4:00 PM

Introduction to IC252

Attendance Requirement: 70%

Other Requirements:

- ▶ Switch off your mobile phone in the classroom/lab.
- ▶ Be punctual to the lecture and lab sessions.

Instructions for weekly theory assignments:

- ▶ Write solutions in A4 sized blank sheets only and in the same order in which problems are given in the assignment.
 - ▶ Separate each solution by a horizontal line from the left end to the right end of the page. For example,
-

Instruction for weekly lab assignments:

- ▶ Keep a separate notebook for the lab related workouts.

Sets

- ▶ The mathematical framework for probability is built around sets.
- ▶ A set is a collection of objects (also called elements).
- ▶ The objects can be anything, e.g., numbers, names
- ▶ Set notation A, B, C, \dots
- ▶ Examples $A = \{1, 2, 3, 4, 5\}$, $B = \{\alpha, \beta, \gamma, \delta\}$
- ▶ The empty set is the smallest set containing no elements.
- ▶ Denoted \emptyset or $\{\}$. Note that, $\emptyset \neq \{\emptyset\}$
empty set *the set containing the empty set as an element.*
- ▶ Subset relation: A is a subset of B , denoted $A \subseteq B$, if every element of A is also an element of B .

Sets

- ▶ Union of sets A, B , denoted $A \cup B$, is the set of all objects that are in A or B .
- ▶ Intersection of sets A, B , denoted $A \cap B$, is the set of all objects that are in both A and B
- ▶ A and B are disjoint sets if $A \cap B = \emptyset$.
- ▶ In many applications, all the sets we're working with are subsets of some set S .
- ▶ Complement of a set A , denoted A^c , is the set of all objects in S that are not in A .
- ▶ DeMorgan's law: $(A \cap B)^c = A^c \cup B^c$ and $(A \cup B)^c = A^c \cap B^c$

Sets

- If A is a finite set, we write $|A|$ for the number of elements in A , which is called its size or cardinality.
- Note that $|A \cup B| = |A| + |B| - |A \cap B|$ (Inclusion-Exclusion Formula for 2 sets).
- Cartesian product of sets A, B is defined as

$$A \times B = \{(a, b) : a \in A, b \in B\}.$$

called tuple (or ordered pair)

- Example

$$\begin{aligned} A &= \{\alpha, \beta\} \\ B &= \{1, 2, 3\} \end{aligned} \Rightarrow A \times B = \{(\alpha, 1), (\alpha, 2), (\alpha, 3), (\beta, 1), (\beta, 2), (\beta, 3)\}.$$

- The above operations results can be generalized for n sets.

Outcomes, sample spaces and events

- ▶ Imagine that an experiment is performed, resulting in one out of a set of possible outcomes.
- ▶ The sample space S of an experiment is the set of all possible outcomes of the experiment.
- ▶ An event A is a subset of the sample space S .
- ▶ We say that A occurred if the actual outcome is in A .
- ▶ The sample space of an experiment (i.e., the cardinality of S) can be finite, countably infinite, or uncountably infinite (advance topic)
- ▶ Example: Experiment of tossing one coin.

sample space : $S = \{H, T\}$

outcomes : H, T

Events : $\emptyset, \{H\}, \{T\}, \{H, T\}$

Outcomes, sample spaces and events

- ▶ Example (Coin flips). A coin is flipped 10 times. Writing Heads as H and Tails as T , a possible outcome is $HHHTHHTTHT$, and the sample space is the set of all possible strings of length 10 of H 's and T 's. We can encode H as 1 and T as 0, so that an outcome is a sequence (s_1, \dots, s_{10}) with $s_j \in \{0, 1\}$, and the sample space is the set of all such sequences. Now let's look at some events:
- ▶ Let A_1 be the event that the first flip is Heads.

$$A_1 = \{(1, s_2, \dots, s_{10}) : s_j \in \{0, 1\}, \text{ for } 2 \leq j \leq 10\}$$

Similarly, define A_j as the event that j^{th} flip is H for $j \in \{2, \dots, 10\}$.

Outcomes, sample spaces and events

- Let B be the event that at least one flip was Heads. Write B in terms of A_j 's:

$$B = \bigcup_{j=1}^{10} A_j.$$

- Let C be the event that all the flips were Heads. Write C in terms of A_j 's:

$$C = \bigcap_{j=1}^{10} A_j.$$

- Let D be the event that there were at least two consecutive Heads. Write D in terms of A_j 's:

$$D = \bigcup_{j=1}^9 (A_j \wedge A_{j+1}).$$

Naive probability

- ▶ The earliest definition of the probability of an event was to count the number of ways the event could happen and divide by the total number of possible outcomes for the experiment.
- ▶ Let A be an event for an experiment with a finite sample space S . The **naive probability** of A is

$$P(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to } A}{\text{total number of outcomes in } S}.$$

- ▶ The naive definition requires S to be finite, and the outcomes to be equally likely.
- ▶ Reading: Blitzstein and Hwang, *Introduction to Probability* (2015), CRC Press.

Lecture 2: Counting - Part I

Satyajit Thakor
IIT Mandi

18 February, 2020

Naive probability (cont.)

- ▶ What is the naive probability of the event A^c ?

$$\begin{aligned} P(A^c) &= \frac{|A^c|}{|S|} \\ &= \frac{|S| - |A|}{|S|} \\ &= 1 - \frac{|A|}{|S|} \\ &= 1 - P(A). \end{aligned}$$

- ▶ We will see later that this is true even for axiomatic (generalized) definition of probability.
- ▶ Assume we want to find $P(A)$. Sometimes, it is easier to find $P(A^c)$ (i.e., finding $|A^c|$ for naive probability). Then, from $P(A^c)$ we can compute $P(A)$.

Counting

- ▶ Calculating the naive probability of an event A involves *counting* the number of outcomes in A and the number of outcomes in the sample space S .
- ▶ Theorem (Multiplication rule). Suppose that Experiment A has a possible outcomes, and for each of those outcomes Experiment B has b possible outcomes. Then the compound experiment has ab possible outcomes.
- ▶ Why?

Let A be the sample space for Experiment A and B be the sample space for Experiment B. Then, there are $|A \times B| = ab$ outcomes of the compound experiment consisting sub-experiments A and B, where

$$A \times B = \{(i, j) : i \in A, j \in B\}.$$

Counting

- ▶ Example: Suppose that 10 people are running a race. Assume that ties are not possible and that all 10 will complete the race, so there will be well-defined first place, second place, and third place winners. How many possibilities are there for the first, second, and third place winners?

- A person cannot take 2 or more places.
 - 10 possibilities for the 1st place. { order does not matter.
 - 9 possibilities for the 2nd place. }
 - 8 possibilities for the 3rd place. }
- Total : $10 \cdot 9 \cdot 8 = 720$.

- ▶ Example: How many squares are there in an 8×8 chessboard?

- A square can be identified uniquely by row no. and column no.
- 8 rows and 8 columns \Rightarrow 64 squares.

Counting

- Example: A set with n elements has 2^n subsets. How?

- To form a subset, for each element, we can either choose it or exclude it. i.e., two possibilities for each element.

- There are n total elements.

\Rightarrow the total no. of subsets are $\underbrace{2 \times \dots \times 2}_n = 2^n$

- E.g., $S = \{1, 2, 3\}$. Its subsets are

$\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$.

Counting

- Theorem (Sampling with replacement). Consider n objects and making k choices from them, one at a time with replacement (i.e., choosing a certain object does not preclude it from being chosen again).
Then there are n^k possible outcomes (here order matters).

- Ex: Rolling a dice k times
 $\Rightarrow 6^k$ possible outcomes

flipping a coin k times
 $\Rightarrow 2^k$ possible outcomes

Counting

- ▶ Theorem (Sampling without replacement). Consider n objects and making k choices from them, one at a time without replacement (i.e., choosing a certain object precludes it from being chosen again).

Then there are $n(n - 1) \cdots (n - k + 1)$ possible outcomes for $1 \leq k \leq n$, and 0 possibilities for $k > n$ (here order matters).

- ▶ By convention, $n(n - 1) \cdots (n - k + 1) = n$ for $k = 1$.
- ▶ Also, note that for $k = n$, $n(n - 1) \cdots (n - k + 1) = n!$. This is the number of ways n objects can be permuted. A permutation of n objects is an arrangement of them in some order, e.g., $(3, 5, 1, 2, 4)$ is a permutation of the objects in $\{1, 2, 3, 4, 5\}$.

- Ex : Refer the "race" problem with 10 participants.

Counting

- ▶ Example (Birthday problem). There are k people in a room. Assume each person's birthday is equally likely to be any of the 365 days of the year (we exclude February 29), and that people's birthdays are independent (we will define independence formally later, but intuitively it means that knowing some people's birthdays gives us no information about other people's birthdays; this would not hold if, e.g., we knew that two of the people were twins).

What is the probability that at least one pair of people in the group have the same birthday?

- There are 365^k ways to assign b'days to k people, i.e., $|S| = 365^k$.
- We are interested in no. of ways to assign b'days s.t. there are at least two people sharing a b'day.

Counting

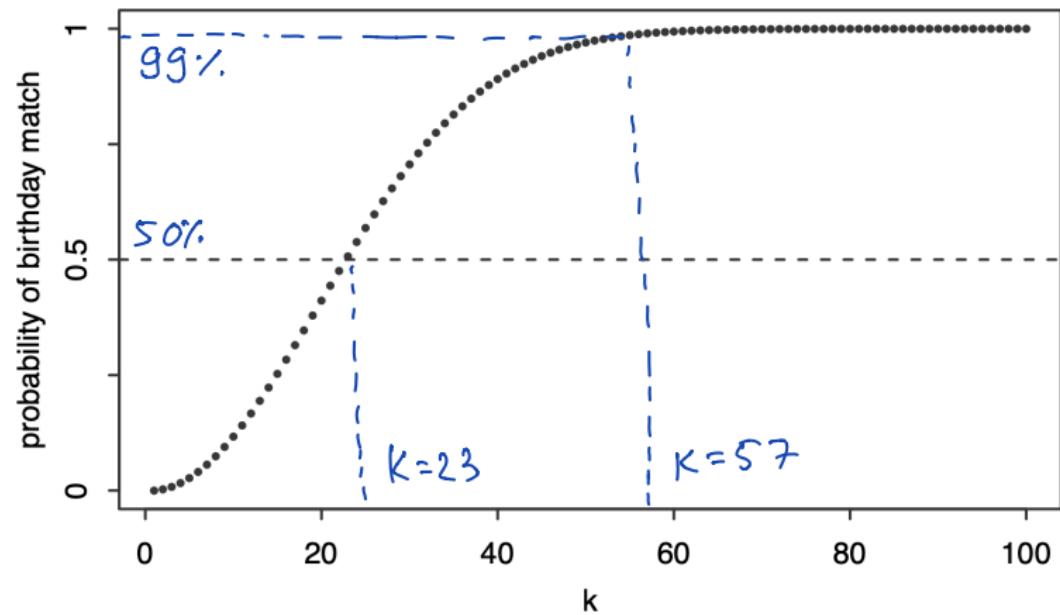
- It is easier to find the size of the complement event, i.e., no. of ways to assign b'day to K people s.t. no two people share a b'day.
- Apply sampling without replacement:
$$365 \cdot 364 \cdot \dots \cdot (365 - k + 1), \quad k \leq 365$$

$$\Rightarrow P(\text{no b'day match}) = \frac{365 \cdot 364 \cdot \dots \cdot (365 - k + 1)}{365^k}$$

$$\Rightarrow P(\text{at least 1 b'day match})$$

$$= 1 - \frac{365 \cdot 364 \cdot \dots \cdot (365 - k + 1)}{365^k}.$$

Counting Note that for $K \geq 366$,
 $P(\text{at least 1 match}) = 1.$



Counting

- Binomial coefficient $\binom{n}{k}$, read as “ n choose k ”, is the number of subsets of size k for a set of size n .
- Theorem:

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{(n-k)!k!}, \text{ for } k \leq n \quad \text{--- (1)}$$
$$= 0, \text{ for } k > n. \quad \text{--- (2)}$$

Proof: There are $n(n-1)\cdots(n-k+1)$ ways to make ordered choices of k elements without replacement.

- This overcounts each subset of interest by $k!$ (no. of permutations)
- Hence (1) follows. For (2), no subset exist of size $k > n$.

Counting

- ▶ Example: In a club with n people,
- ▶ How many ways to choose a president, vice president, and treasurer?

$$n(n-1)(n-2)$$

- ▶ How many ways to choose 3 officers without predetermined titles?

$$\frac{n(n-1)(n-2)}{3!}$$

Counting

► Theorem:

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

Proof 1:
(Algebraic)

$$\binom{n}{k} = \frac{n!}{(n-k)! k!} = \frac{n!}{k! (n-k)!} = \binom{n}{n-k}.$$

Proof 2:
(Intuitive) choosing a committee of size k from n people is the same as choosing $\binom{n}{k}$ people not on the committee.

(Specifying who is on the committee
also determines who is not on the
committee.)

Counting

- ▶ How many ways are there to permute the letters in the word LALALAAA?

- Out of 8 positions, choose 5 positions where A goes.
- OR: choose 3 positions where L goes.
 $\Rightarrow \binom{8}{5} = \binom{8}{3} = \frac{8 \cdot 7 \cdot 6}{3!} = 56.$

- ▶ Preliminary reading: Blitzstein and Hwang, *Introduction to Probability* (2015), CRC Press.

Lecture 3: Counting - Part II & Axiomatic Probability

Satyajit Thakor
IIT Mandi

24 February, 2020

Counting

- Example: A 5-card hand is dealt from a standard, well-shuffled 52-card deck. The hand is called a full house in poker if it consists of three cards of some rank and two cards of another rank, e.g., three 7's and two 10's (in any order). What is the probability of a full house.

- $|S| = \binom{52}{5}$.
- 13 choices for what rank we have of 3 cards for a full house. Fixing some rank i , there are $\binom{4}{3}$ ways to choose which 3 cards of rank i we have.
- 12 choices of what rank we have of 2 cards for a full house. Fixing some rank j , there are $\binom{4}{2}$ ways to choose 2 cards for a full house.
$$\Rightarrow P(\text{full house}) = \frac{13 \binom{4}{3} 12 \binom{4}{2}}{\binom{52}{5}} = \frac{3744}{2598960} \approx 0.00144.$$

Counting

- The factorial function $n!$ grows extremely quickly as n grows. A famous, useful approximation for factorials is Stirling's formula:

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

- The ratio of the two sides converges to 1 as $n \rightarrow \infty$, i.e., the "error" in approximation reduces as n grows.
- Example (approximating the number of permutations): Suppose that we want to compute the number of tuples of 20 objects by choose objects from a set of 70 objects, i.e., $n!/(n - k)! = 70!/50!$. The approximation from Stirling's formula is

$$\frac{70!}{50!} \approx \frac{\sqrt{140\pi}(70/e)^{70}}{\sqrt{100\pi}(50/e)^{50}} = 3.940 \times 10^{35}.$$

The exact calculation yields 3.938×10^{35} .

Counting

- Suppose that 20 members of an organization are to be divided into three committees A, B, and C in such a way that each of the committees A and B is to have eight members and committee C is to have four members. Determine the number of different ways in which members can be assigned to these committees. (we have discussed a similar problem in the previous lecture)

- $\binom{20}{8}$ ways to choose members for Committee A.

- $\binom{12}{8}$ " " " " " " B.

- $\binom{4}{4} = 1$ way " " " " " " C.

$$\Rightarrow \text{The total no. of ways} = \binom{20}{8} \binom{12}{8} = \frac{20!}{8! 12!} \cancel{\frac{12!}{8! 4!}} \\ = \frac{20!}{8! 8! 4!} = 62355150.$$

Counting

- **Multinomial coefficient:** Suppose that n distinct elements are to be divided into k different groups ($k \geq 2$) in such a way that, for $j = 1, \dots, k$, the j th group contains exactly n_j elements, where

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

Then, the number of different ways in which the n elements can be divided into the k groups is

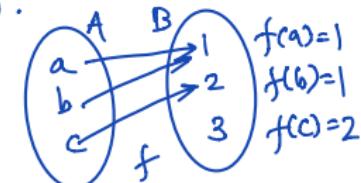
$$\frac{\binom{n}{n_1} \binom{n-n_1}{n_2} \binom{n-n_1-n_2}{n_3} \dots \binom{n-n_1-\dots-n_{k-2}}{n_{k-1}}}{\frac{n!}{n_1! n_2! \dots n_k!}} = \frac{(n-n_1-n_2-\dots-n_{k-2})!}{(n-n_1)! n_1! \cdot (n-n_1-n_2)! n_2! \cdot (n-n_1-n_2-n_3)! n_3! \cdot \dots \cdot (n-\sum_{i=1}^{k-2} n_i)! n_{k-1}!}$$

Axiomatic probability

- ▶ Previously, we discussed naive probability. Now we will discuss the most general definition of probability.
- ▶ The definition is **axiomatic**, i.e., defined by a set of rules.
- ▶ First, question: what is probability (mathematically)?
A function.
- ▶ What is a function?

A function f from a set A to a set B maps each element of A to some element of B .

$\rightarrow f: A \rightarrow B$
function ↗ domain ↙ codomain



Recall: range/image $\triangleq \{b \in B : b = f(a) \text{ for some } a \in A\}$.
range = codomain

injective (one-to-one), $\underbrace{f(a) \neq f(b), \forall a, b \in A, a \neq b}$, surjective, bijection
inj & surj

Axiomatic probability

- A probability space consists of a sample space S and a probability function P which takes an event $A \subseteq S$ as input and returns $P(A)$, a real number between 0 and 1, as output, i.e.,

$P : \{A \subseteq S : A \text{ is an event}\} \longrightarrow [0, 1].$

$\xrightarrow{\text{probability function}}$ $\underbrace{\text{set of all events}}_{\triangleq \{x \in \mathbb{R} : 0 \leq x \leq 1\}}$
 $\text{i.e. set of real nos. } x \text{ s.t. } 0 \leq x \leq 1.$

The function P must satisfy the following axioms:

Axiom 1: $P(\emptyset) = 0, P(S) = 1.$

Axiom 2: If A_1, A_2, \dots are disjoint events, then

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

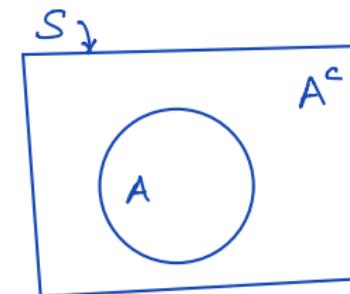
- Events are disjoint means that they are mutually exclusive, i.e., $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Properties of probability function

- ▶ Several properties of the probability function can be derived by its axiomatic definition.
- ▶ **Property 1:** $P(A^c) = 1 - P(A)$ (Recall: A similar property was discussed for naive probability)

Proof:

$$\begin{aligned} P(S) &= P(A \cup A^c) \\ &= P(A) + P(A^c) \quad (\because \text{Axiom 2}) \\ &= 1 + P(A^c) \quad (\because \text{Axiom 1}) \\ \Rightarrow 1 &= P(A) + P(A^c) \\ \Rightarrow P(A^c) &= 1 - P(A). \end{aligned}$$



Properties of probability function

- **Property 2:** If $A \subseteq B$ then $P(A) \leq P(B)$.

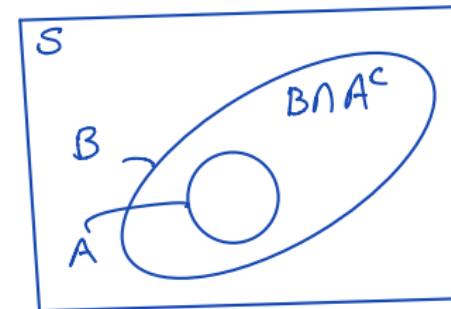
$$P(B) = P(A \cup \underbrace{B \cap A^c}_{\text{also denoted } B \setminus A})$$

set subtraction

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

$(\because \text{Axiom 2})$
 $(\because \text{Axiom 1})$

$$P(B) \geq P(A) + 0$$



Properties of probability function

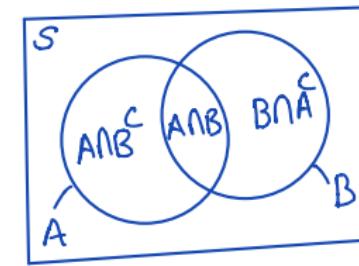
- **Property 3:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

$$\begin{aligned}P(A \cup B) &= P(A \cup (B \cap A^c)) \\&= P(A) + P(B \cap A^c) \quad (\because \text{Axiom 2}) \quad \text{--- 1}\end{aligned}$$

$$\begin{aligned}P(B) &= P((A \cap B) \cup (A^c \cap B)) \\&= P(A \cap B) + P(A^c \cap B) \quad (\because \text{Axiom 2}) \quad \text{--- 2}\end{aligned}$$

From ① and ②,

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



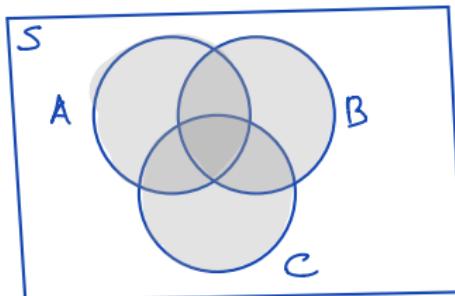
- **Property 3** is the inclusion-exclusion formula (IEF) for the probability function.
- Recall (IEF for cardinality): $|A \cup B| = |A| + |B| - |A \cap B|$.
- Note that, cardinality too is a function (like probability). *same domains*

Properties of probability function

- IEF for 3 events:

$$\begin{aligned} P(A \cup B \cup C) = & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ & + P(A \cap B \cap C). \end{aligned}$$

Proof : (homework.
(obtain an algebraic proof))



Properties of probability function

- ▶ Generalization: The IEF holds for the probability function involving n events:
- ▶ **Theorem:** For events A_1, \dots, A_n ,

$$\begin{aligned} P\left(\bigcup_{i=1}^n A_i\right) &= \sum_{i=1}^n P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) - \\ &\quad \dots + (-1)^{n+1} P(A_1 \cap \dots \cap A_n)] \\ &= \sum_{k=1}^n \left((-1)^{k-1} \sum_{I \subseteq \{1, 2, \dots, n\}: |I|=k} P(\cap_{i \in I} A_i) \right). \end{aligned}$$

- ▶ Can be proved using mathematical induction.

Lecture 4:
Counting - Part III
&
Conditional Probability - Part I

Satyajit Thakor
IIT Mandi

25 February, 2020

Counting

- de Montmort's matching problem: Consider a well-shuffled deck of n cards, labelled 1 through n . You flip over the cards one by one, saying the numbers 1 through n as you do so. You win the game if, at some point, the number you say aloud is the same as the number on the card being flipped over (for example, if the 7th card in the deck has the label 7). What is the probability of winning?
- What is your guess: How the probability will grow as $n \rightarrow \infty$?
- Hint: To solve the problem employ the IEF.

Sol: Let A_i be the event that i th card has the no.
 i written on it.

We want to find $\Pr(A_1 \cup A_2 \cup \dots \cup A_n)$
Pr. that there is at least 1 card s.t. no. said = no. written
= Pr. of winning the game.

Counting

$$- P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}$$

$$- P(A_i \wedge A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$$

$$- P(A_i \wedge A_j \wedge A_l) = \frac{1}{n(n-1)(n-2)}$$

$$- P\left(\bigcap_{\substack{i \in I : |I|=k \\ I \subseteq \{1, \dots, n\}}} A_i\right) = \frac{1}{n(n-1)(n-2) \dots (n-k+1)}$$

IEF:

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{\substack{i=1 \\ n \text{ possibilities}}}^n P(A_i) - \sum_{\substack{\{i,j\} \subseteq \{1, \dots, n\} \\ (2)}} P(A_i \wedge A_j)$$
$$+ \sum_{\substack{(3) \\ \{i,j,k\} \subseteq \{1, \dots, n\}}} P(A_i \wedge A_j \wedge A_k) - \dots + (-1)^{n+1} P\left(\bigcap_{j=1}^n A_j\right)$$

Counting

$$= \frac{n}{n} - \frac{\binom{n}{2}}{n(n-1)} + \frac{\binom{n}{3}}{n(n-1)(n-2)} - \dots + (-1)^{n+1} \frac{1}{n!}$$

$$= 1 - \frac{1}{2!} + \frac{1}{3!} - \dots + \frac{(-1)^{n+1}}{n!}$$

But, the Taylor series expansion for e^x is:

$$\begin{aligned}\frac{1}{e} &= 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \dots \\ &= 1 - \underbrace{\left(1 - \frac{1}{2!} + \frac{1}{3!} - \dots \right)}_{P(\bigcup_{i=1}^{\infty} A_i)}\end{aligned}$$

$$\Rightarrow \text{as } n \rightarrow \infty \quad P(\text{winning}) = 1 - \frac{1}{e} \approx 0.63.$$

Conditional probability

- ▶ Roughly speaking, conditional probability is the concept that addresses this fundamental question: how should we update our beliefs in light of the evidence we observe?
- ▶ Conditional probability is essential for reasoning in many fields, e.g., scientific, legal etc.

event of interest

- ▶ Example 1: What is the probability of rain? What is the probability of rain given that the sky is clear?
event is evidence/observation
- ▶ Example 2: What is the probability that John has stolen a car? What is the probability that John has stolen a car given that John has been convicted stealing 5 cars in the past and that car's tire marks are found at John's property?
two evidence/observations.

Conditional probability

- If A and B are events with $P(B) > 0$, then the conditional probability of A given B , denoted by $P(A|B)$, is defined as

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

- $P(A|B)$: the probability of the event A provided that the event B (an evidence) has occurred.
- $P(A)$ is called the prior probability of A and $P(A|B)$ is called the posterior probability of A .
- “prior” means before updating based on the evidence, and “posterior” means after updating based on the evidence.
- For any event A , $P(A|A) = P(A \cap A)/P(A) = 1$. That is, upon observing that A has occurred, our updated probability for A is 1.
makes sense!

Conditional probability

- Example: A standard deck of cards is shuffled well. Two cards are drawn randomly, one at a time without replacement. Let A be the event that the first card is a heart, and B be the event that the second card is red. Find $P(A|B)$ and $P(B|A)$.

Sol: $P(A \cap B) = \frac{13}{52} \cdot \frac{25}{51} = \frac{25}{204}$.

$$P(A) = \frac{13}{52} = \frac{1}{4}$$

$$P(B) = \frac{26}{52} = \frac{1}{2}$$

- Rather than thinking as cards drawn 1st and 2nd,
you may think as cards drawn (simultaneously)
by right and left hand. (order does not matter)

Conditional probability
- Hence we can directly see that $P(B) = \frac{26}{52} = \frac{1}{2}$.

$$\Rightarrow P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{25/204}{\frac{1}{2}} = \frac{25}{102},$$

$$\text{and } P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{25/204}{\frac{25}{51}} = \frac{25}{51}.$$

Observe: $P(A|B) \neq P(B|A)$, conditioning is not symmetrical.

① $P(A|B) \neq P(B|A)$, (direct argument)

② Easy to see why $P(B|A) = \frac{25}{51}$ (direct argument)

③ $P(A|B)$: not straightforward to see.
"red has occurred" does not help much
(more interpretation in later lectures)

Conditional probability
A statement/problem which leads to a contradiction.

- Two children paradox: Martin Gardner posed the following puzzle in the 1950s:
 - (1) Mr. Jones has two children. The older child is a girl. What is the probability that both children are girls?
 - (2) Mr. Smith has two children. At least one of them is a boy. What is the probability that both children are boys?
- What is your guess: Should the answer to (1) and (2) be the same? Why?

$$(1) P(\text{GG} | \text{elder is G}) = \frac{P(\text{GG}, \text{elder is G})}{P(\text{elder is G})} = \frac{1/4}{1/2} = \frac{1}{2}$$

$$(2) P(\text{BB} | \text{at least 1 B}) = \frac{P(\text{BB}, \text{at least 1 B})}{P(\text{at least 1 B})} = \frac{1/4}{3/4} = \frac{1}{3}$$

Conditional probability

- Contradiction occurs since it is not clear the event "at least 1 B" is (or should be) generated.
- Alternative way to generate "at least 1 B":

Older child	Younger child	$P(\text{this family})$	$P(\text{AL1B given this family})$	$P(\text{AL1B and this family})$
A	A	$\frac{1}{4}$	0	0
A	B	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$
B	A	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$
B	B	$\frac{1}{4}$	1	$\frac{1}{4}$

if you let 1 then $P(\text{AL1A given this family})$
will be 0. (Which may seem absurd to you.)

You need to understand the law of total prop. to understand this

$\Rightarrow P(\text{AL1B}) = \frac{1}{8} + \frac{1}{8} + \frac{1}{4} = \frac{1}{2}$

$\Rightarrow \frac{P(\text{AL1B} \cap \text{BB})}{P(\text{AL1B})} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$

Lecture 5: Conditional Probability - Part II

Satyajit Thakor
IIT Mandi

28 February, 2020

Conditional probability

- ▶ By the definition of conditional probability,

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

- ▶ For three events A_1, A_2, A_3 with positive probabilities,

$$\begin{aligned} P(A_1, A_2, A_3) &= P(A_1|A_2, A_3)P(A_2, A_3) \quad ; \leftrightarrow \cap \\ &= P(A_1|A_2, A_3)P(A_2|A_3)P(A_3) \end{aligned}$$

- ▶ Similarly, $P(A_1, A_2, A_3) = P(A_2|A_1, A_3)P(A_1|A_3)P(A_3)$ and in fact, we can have 6 distinct expressions by permuting A_1, A_2, A_3 .

- “,” is used to denote “ \cap ”.

- Writing the probability of intersection of events as product of conditional probabilities.

Conditional probability

- **Theorem:** For events A_1, \dots, A_n with $P(A_1, A_2, \dots, A_{n-1}) > 0$,

$$P(A_1, A_2, \dots, A_n) = P(A_1)P(A_2|A_1) \dots P(A_n|A_1, \dots, A_{n-1}).$$

- In fact, we have $n!$ theorems by permuting A_1, A_2, \dots, A_n .
- Theorem (Bayes' rule):

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

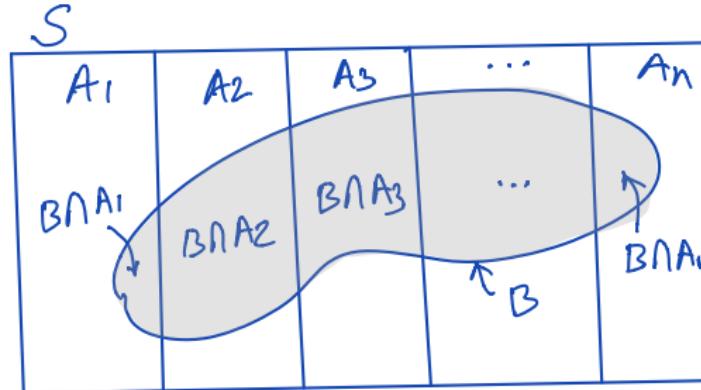
- The odds of an event A are $\text{odds}(A) = P(A)/P(A^c)$.
- E.g., if $P(A) = 2/3$ then the odds of A are 2 to 1.
- Note that

$$\begin{aligned} P(A) &= \text{odds}(A)(1 - P(A)) \\ \Rightarrow P(A) &= \frac{\text{odds}(A)}{1 + \text{odds}(A)}. \end{aligned}$$

Conditional probability

- A_1, \dots, A_n is a partition of set S if $A_i \cap A_j = \emptyset, \forall i \neq j$ (i.e., disjoint sets) and $A_1 \cup \dots \cup A_n = S$.
- Theorem (Law of total probability LOTP): Let A_1, \dots, A_n be a partition of the sample space S with $P(A_i) > 0$ for all i . Then

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



Conditional probability

$$\begin{aligned} p(B) &= P(B \cap S) \\ &= P\left(B \cap (A_1 \cup A_2 \cup \dots \cup A_n)\right) \\ &= P\left(\underbrace{(B \cap A_1)}_{\text{disjoint events}} \cup \underbrace{(B \cap A_2)}_{\text{}} \cup \dots \cup \underbrace{(B \cap A_n)}_{\text{}}\right) \quad (\because \text{distributive property of } \cap \text{ over } \cup) \\ &= P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_n) \\ &= P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_n) \quad (\because \text{Axiom 2}) \\ &= P(B|A_1) \cdot P(A_1) + \dots + P(B|A_n) \cdot P(A_n) \\ &= \sum_{i=1}^n P(B|A_i) \cdot P(A_i) \end{aligned}$$

Conditional probability

- ▶ Example: You have one fair coin, and one biased coin which lands Heads with probability $3/4$. You pick one of the coins at random and flip it three times. It lands Heads all three times. Given this information, what is the probability that the coin you picked is the fair one?

$A =$ the event that the chosen coin land 3 Heads.
 $F =$ the event that the chosen coin is the fair coin.

We want to find $P(F|A)$.

- If is not easy to find $P(F|A)$ directly.
- If is easier to find $P(A|F)$, $P(A|F^C)$.

Conditional probability

By Bayes' rule,

$$P(F|A) = \frac{P(A|F) \cdot P(F)}{P(A)} = \frac{P(A|F) \cdot P(F)}{\underbrace{P(A|F) \cdot P(F) + P(A|F^c) P(F^c)}_{(\because \text{ LOTP})}}$$
$$= \frac{\left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)}{\left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right) + \left(\frac{3}{4}\right)^3 \cdot \frac{1}{2}}$$

$$\approx 0.23$$

I.e., probability that the coin is fair given that the first 3 are Heads is less than the coin is biased:

$$P(F^c|A) = 1 - P(F|A)$$

$$\approx 0.77. \quad (\because P(S|A) = P(F|A) + P(F^c|A))$$

Conditional probability

- Theorem: Conditional probabilities given an “evidence” event E are all probabilities.

Proof: We want to prove that $P(\cdot | E)$ is a (valid) probability function.
I.e., check whether Axioms 1,2 are true for $P(\cdot | E)$.

$$\textcircled{1} \quad P(\emptyset | E) = \frac{P(\emptyset \wedge E)}{P(E)} = \frac{P(\emptyset)}{P(E)} = 0$$

$$P(S | E) = \frac{P(S \wedge E)}{P(E)} = \frac{P(E)}{P(E)} = 1$$

Conditional probability
② Let A_1, A_2, \dots, A_n be disjoint events. Then,

$$\begin{aligned} & P((A_1 \cup A_2 \cup \dots \cup A_n) | E) \\ &= \frac{P((A_1 \cup A_2 \cup \dots \cup A_n) \cap E)}{P(E)} \\ &= \frac{P((A_1 \cap E) \cup (A_2 \cap E) \cup \dots \cup (A_n \cap E))}{P(E)} \\ &= \sum_{i=1}^n \frac{P(A_i \cap E)}{P(E)} \\ &= \sum_{i=1}^n P(A_i | E) \quad \text{i.e., } P(\cdot | E) \text{ also satisfies Axiom 2.} \end{aligned}$$

Conditional probability

- Theorem (Bayes' rule with extra conditioning): If $P(A, E) > 0$ and $P(B, E) > 0$ then

$$P(A|B, E) = \frac{P(B|A, E)P(A|E)}{P(B|E)}.$$

Proof:
$$\begin{aligned} P(A|B, E) &= \frac{P(A, B, E)}{P(B, E)} \\ &= \frac{P(B|A, E) \cdot P(A|E) \cdot P(E)}{P(B|E) \cdot P(E)} \end{aligned}$$

Conditional probability

- Theorem (LOTP with extra conditioning): Let A_1, \dots, A_n be a partition of S with $P(A_i, E) > 0$ for all i . Then

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

Proof: $P(B|E) = \frac{P(B \cap E \cap S)}{P(E)}$

$$= \frac{P(B \cap E \cap (A_1 \cup \dots \cup A_n))}{P(E)}$$
$$= \sum_{i=1}^n \frac{P(B \cap E \cap A_i)}{P(E)}$$
$$= \sum_{i=1}^n P(B|A_i, E) \cdot \frac{P(A_i, E)}{P(E)}$$

Conditional probability

- Example: Continuing with the “coin” example, suppose that we have now seen our chosen coin land Heads three times. If we toss the coin a fourth time, what is the probability that it will land Heads once more?

$A = \text{event that the chosen coin has 3 H.}$

$B = \text{" " " " " " " 4^{th} H.}$

We want to find $P(B|A)$.

$$P(B|A) = P(B|A, F) \cdot P(F|A) + P(B|A, F^c) \cdot P(F^c|A)$$

$$\approx \frac{1}{2}(0.23) + \frac{3}{4}(1 - 0.23)$$

$$\approx 0.69.$$

Conditional probability assume that it occurred between 2pm-3pm.

- Let G be the event that a certain individual is guilty of a certain robbery. In gathering evidence, it is learned that an event E_1 occurred, and a little later it is also learned that another event E_2 also occurred.

Is it possible that individually, these pieces of evidence increase the chance of guilt (so $P(G|E_1) > P(G)$ and $P(G|E_2) > P(G)$), but together they decrease the chance of guilt (so $P(G|E_1, E_2) < P(G)$)?

Let $E_1 = A$ was in a nearby restaurant between 2pm-3pm.
 $E_2 = " " " " "$.. 3pm-4pm.

- $P(\alpha)$ may be taken as $1/\text{population of the town}$.

- Note that $P(\alpha|E_i) > P(\alpha)$, $i=1, n$.
(near the incident \Rightarrow more prob.)

- But $P(\alpha|E_1, E_2) = 0 < P(\alpha)$.

Lecture 6: Conditional Probability - Part III & Independence

Satyajit Thakor
IIT Mandi

2 March, 2020

Conditional probability

- ▶ Example: A patient named Fred is tested for a disease called conditionitis, a medical condition that afflicts 1% of the population. The test result is positive, i.e., the test claims that Fred has the disease. Let D be the event that Fred has the disease and T be the event that he tests positive. Suppose that the test is “95% accurate”, i.e., $P(T|D) = 0.95$ and $P(T^c|D^c) = 0.95$. The quantity $P(T|D)$ is known as the sensitivity or true positive rate of the test, and $P(T^c|D^c)$ is known as the specificity or true negative rate.
Find the conditional probability that Fred has conditionitis, given the evidence provided by the test result.

- We want to find $P(D|T)$.

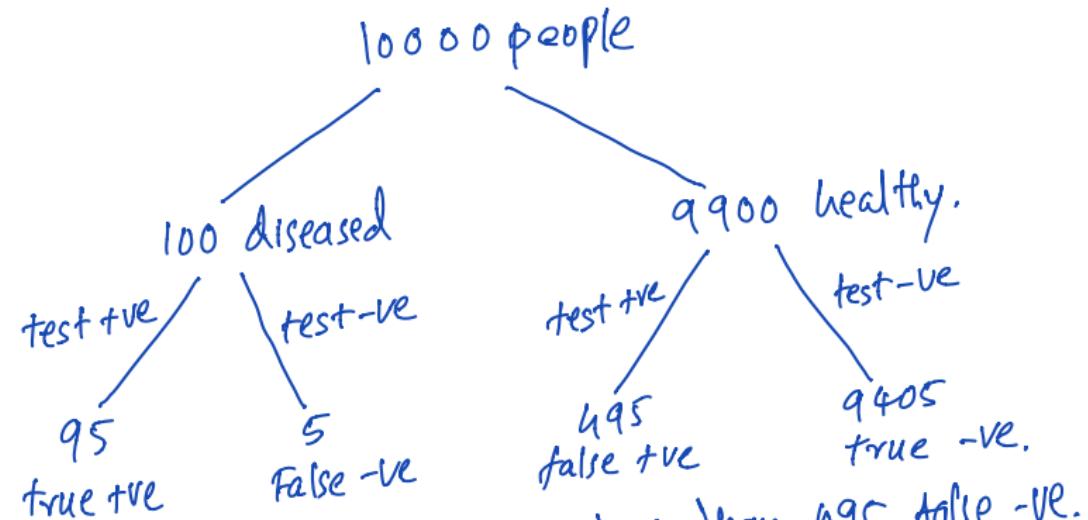
Conditional probability

By Bayes' rule,

$$\begin{aligned} P(D|T) &= \frac{P(T|D) \cdot P(D)}{P(T)} \\ &= \frac{P(T|D) \cdot P(D)}{P(T|D) \cdot P(D) + P(T|D^c) \cdot P(D^c)} \\ &= \frac{0.95 \cdot 0.01}{(0.95 \cdot 0.01) + (0.05 \cdot 0.99)} \end{aligned}$$

- Why such low probability? (≈ 0.16 when the test is 95% accurate)
- Since the disease is rare.

Conditional probability



- Note that, 95 true +ve are far less than 495 false -ve.
- Hence, when tested +ve, the probability of
true +ve is small : $\frac{\text{no. of true +ve's}}{\text{no. of all +ve's}}$

Independence

- ▶ Events A and B are independent if $P(A \cap B) = P(A)P(B)$.
- ▶ If $P(A) > 0$ and $P(B) > 0$, then this is equivalent to

$$\begin{aligned}P(A|B) &= P(A), \\P(B|A) &= P(B).\end{aligned}$$

- ▶ I.e., A and B are independent if learning that B occurred does not change the probabilities for A occurring (and vice versa).
- ▶ Independence is a symmetric relation: if A is independent of B , then B is independent of A .
- ▶ Independence is completely different from disjointness. Disjoint events can be independent only if $P(A) = 0$ or $P(B) = 0$.

$$P(A) > 0, P(B) > 0, A \cap B = \emptyset.$$

$$\Rightarrow P(A \cap B) = 0, P(A)P(B) > 0.$$

Independence

- A card is selected at random from an ordinary deck of 52 playing cards. E is the event that the selected card is an ace and F is the event that it is a spade. Are E and F independent?



$$P(E) = 4/52, \quad P(F) = 13/52$$



$$P(E \cap F) = 1/52 \Rightarrow E \text{ & } F \text{ are independent.}$$

- Suppose that we toss 2 fair dice. Let E denote the event that the sum of the dice is 6 and F denote the event that the first die equals 4. Are E and F independent?

$$P(E) = 5/36, \quad P(F) = 1/6$$

$$P(E \cap F) = P(\text{the outcome is } (4,2))$$

$$= 1/36 \Rightarrow E \text{ & } F \text{ are not independent.} \\ (\text{they are dependent})$$

Independence

- Lemma: If A and B are independent, then A and B^c are independent.

Proof: - If $P(A) = 0$ then, A is independent of

any event.

- Now, assume that $P(A) > 0$. Then,

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

$$= 1 - P(B)$$

$$= P(B^c).$$

- Homework: Show that if A and B are independent, then A^c and B are independent, and A^c and B^c are independent.

Independence

- Events A , B , and C are said to be **independent** if all of the following equalities hold:

$$\left\{ \begin{array}{l} P(A \cap B) = P(A)P(B), \\ P(A \cap C) = P(A)P(C), \\ P(B \cap C) = P(B)P(C), \\ P(A \cap B \cap C) = P(A)P(B)P(C). \end{array} \right.$$

- If the first three conditions hold, then A , B , and C are called **pairwise independent**.

Independence

- ▶ Pairwise independence doesn't imply independence.

Example: Consider two fair, independent coin tosses, and let A be the event that the first is Heads, B the event that the second is Heads, and C the event that both tosses have the same result.

$$\begin{aligned} P(A) = P(B) &= \frac{1}{2} & P(C) &= \frac{1}{2} & \text{Similarly } P(B \cap C) &= \frac{1}{4} \\ P(A \cap C) &= P(C|A) \cdot P(A) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} & & & & \\ P(A \cap B \cap C) &= P(C|A \cap B) \cdot P(A) \cdot P(B) \\ &= 1 \cdot P(A) \cdot P(B) = \frac{1}{4} \neq P(A) \cdot P(B) \cdot P(C) \end{aligned}$$

Or find no. of ways the events $A \cap C$, $A \cap B \cap C$ can occur. = 18.

- ▶ $P(A \cap B \cap C) = P(A)P(B)P(C)$ does not imply pairwise independence.

Example: In the “dice” example, assume that $G = \emptyset$. Then

$P(E \cap F \cap G) = P(E)P(F)P(G) = 0$. However, E and F are not independent.

Independence

- Events A_1, A_2, \dots, A_n are independent if

$$P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j) \quad \text{for all } J \subseteq \{1, 2, \dots, n\}.$$

- Events A_1, A_2, \dots, A_n are pairwise independent if

$$P(A_i \cap A_j) = P(A_i)P(A_j) \quad \text{for all } \{i, j\} \subseteq \{1, 2, \dots, n\}, i \neq j.$$

- Events A and B are conditionally independent given E if

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

- $P(A \cap B) = P(A)P(B)$ does not imply

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

Independence

(i.e., hungry or tired)

- A certain baby cries if and only if she is hungry, tired, or both.
Let C be the event that the baby is crying, H be the event that she is hungry, and T be the event that she is tired. Let $P(C) = c$, $P(H) = h$, and $P(T) = t$, where none of c, h, t are equal to 0 or 1. Let H and T be independent.
 - Find c , in terms of h and t .
 - Are H and T conditionally independent given C ?

$$\begin{aligned}(a) \quad P(C) &= P(H \cup T) \\ &= P(H) + P(T) - P(H \cap T) \quad (\because \text{IEF}) \\ &= h + t - \underbrace{ht}_{(\because H \text{ & } T \text{ are independent})}\end{aligned}$$

Independence

(b) check: $P(H \cap T | C) = P(H|C) \cdot P(T|C)$

$$P(H|C) = \frac{\overbrace{P(C|H)}^1 \cdot P(H)}{P(C)} = \frac{1}{c}$$

$$P(T|C) = \frac{\overbrace{P(C|T)}^1 \cdot P(T)}{P(C)} = \frac{t}{c}$$

$$P(H \cap T | C) = \frac{\overbrace{P(C|H,T)}^1 \cdot P(H,T)}{P(C)} = \frac{ht}{c}$$

$$\Rightarrow P(H,T | C) = \frac{ht}{c} \neq P(H|C)P(T|C) = \frac{ht}{c^2}$$

i.e., not conditionally independent.

Lecture 7: Conditional Probability - Part IV & Discrete Random Variables

Satyajit Thakor
IIT Mandi

3 March, 2020

Random Variables

- ▶ Usually, we may not be interested in the outcome of an experiment itself but rather in some numerical function of the outcome or an event.
- ▶ The idea of random variable enables us to put aside the complex structure of the sample space and its events by assigning a real number to the elements of the sample space.
- ▶ Thus we can simply work with numbers rather than complexly defined outcomes and events.
- ▶ Given an experiment with sample space S , a random variable (r.v.) is a function from the sample space S to the real numbers \mathbb{R} .

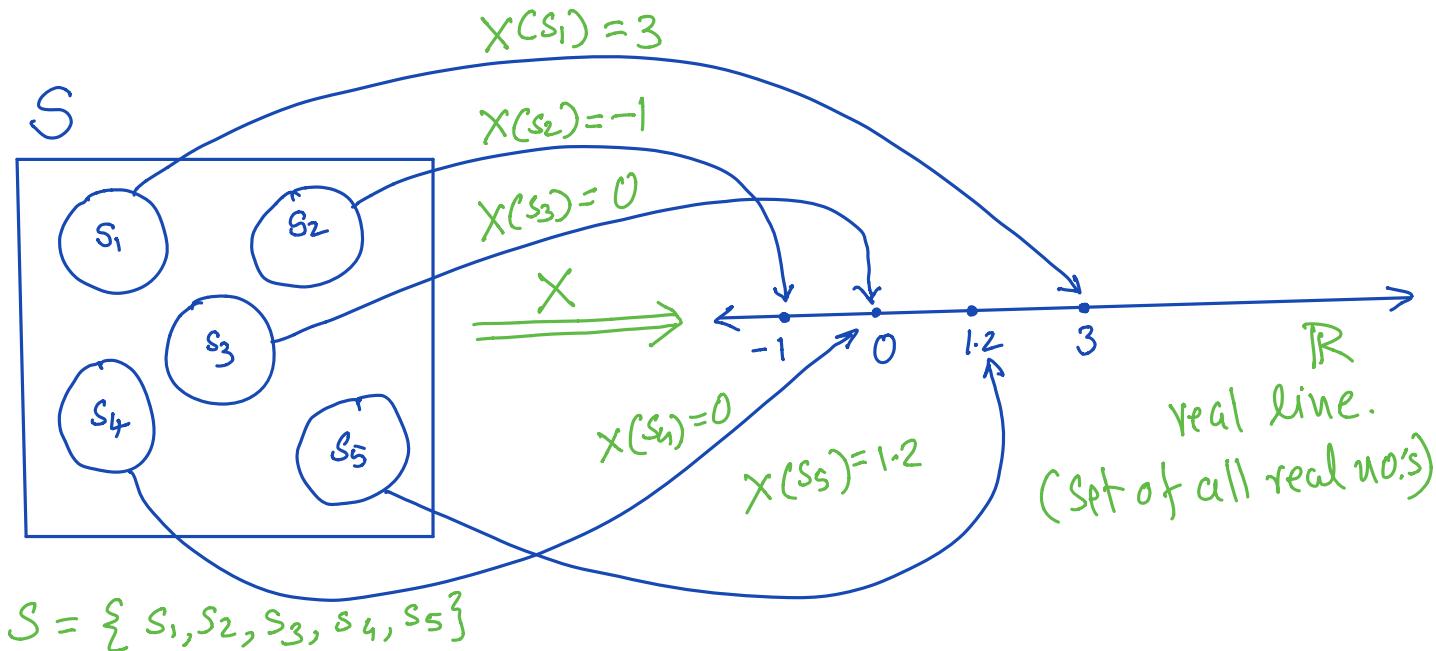
An r.v. is a function : $S \rightarrow \mathbb{R}$.

Random Variables

- It is common to denote random variables by capital letters. For example,

$$X : S \rightarrow \mathbb{R}$$

denotes a random variable X and to each $s \in S$ it assigns a numerical value (real number) $X(s)$.



Random Variables

- ▶ Example: Consider an experiment where we toss a fair coin twice. The sample space consists of four possible outcomes: $S = \{HH, HT, TH, TT\}$. Here are some random variables on this space. Each r.v. is a numerical summary of some aspect of the experiment.
- ▶ Let X be the number of Heads. Then,

$$X(HH) = 2, X(HT) = X(TH) = 1, X(TT) = 0.$$

- ▶ Let Y be the number of Tails. Then, $y(HH) = 0, y(HT) = y(TH) = 1, y(TT) = 2$. I.e., $\underbrace{y(s) = X(s) - 2}$.
the r.v. Y is a function of the r.v. X .

- ▶ Let I be ~~be~~ 1 if the first toss is Heads and 0 otherwise.

$$I(HH) = I(HT) = 1$$

$$I(TH) = I(TT) = 0$$

Such an r.v. is called an "indicator" r.v. since it indicates whether an event has occurred (1) or not (0).

Random Variables

- ▶ Note that, a random variable is neither random nor a variable!
It is a function.
- ▶ The source of randomness is the choice of an outcome (from S) of the experiment according to the probability function P .
- ▶ A random variable X is said to be discrete if there is a finite list of values a_1, a_2, \dots, a_n or an infinite list of values a_1, a_2, \dots such that $P(X = a_j \text{ for some } j) = 1$. i.e., $P(X=a_1 \text{ or } X=a_2 \text{ or } \dots) = 1$.
- ▶ If X is a discrete r.v., then the finite or countably infinite set of values x such that $P(X = x) > 0$ is called the support of X .
- ▶ Example: In the “Two-coin” experiment

- $P(X=0 \text{ or } X=1 \text{ or } X=2) = 1$

- The set $\{0, 1, 2\}$ is finite. Hence X is a discrete random variable.

Random Variables

- ▶ Given a random variable, we would like to be able to describe its behaviour using the language of probability.
- ▶ For example, we might want to answer questions about the probability that the r.v. will fall into a given range: if L is the lifetime earnings of a randomly chosen U.S. college graduate, what is the probability that L exceeds a million dollars?
I.e., find $P(L > 1000000)$.
- ▶ The distribution of a random variable provides the answers to such questions.
- ▶ There are several equivalent ways to express the distribution of an r.v., e.g., cumulative distribution function, probability mass function.
- ▶ For a discrete r.v., the most natural way to do so is with a probability mass function.

Random Variables

i.e., $P_X : \mathbb{R} \rightarrow [0,1]$.

- The probability mass function (PMF) of a discrete r.v. X is the function p_X given by $p_X(x) = P(X = x)$. Note that this is positive if x is in the support of X , and 0 otherwise.

- Note that the event that X takes the value x , i.e., $\{X = x\}$
the subscript X is used to specify that the PMF is for X (not some other r.v. Y)
is equivalent to the event that “ s occurred where $X(s) = x$ ”, i.e.,
 $\{s \in S : X(s) = x\}$.
e.g., $\{HT, TH\}$ ($\because X(HT) = 1$ and $X(TH) = 1$)

- But the notation $\{X = x\}$ is shorter, more convenient and does not involve reference to the sample space.
- Example: In the “Two-coins” example $\{X = 1\}$ refers to $\{HT, TH\}$, i.e., the event that “the outcome is HT or TH ”.

Random Variables

- ▶ Example: Find the PMFs of all the random variables in the “Two-coins” example considering fair coins.
- ▶ X : the number of Heads.

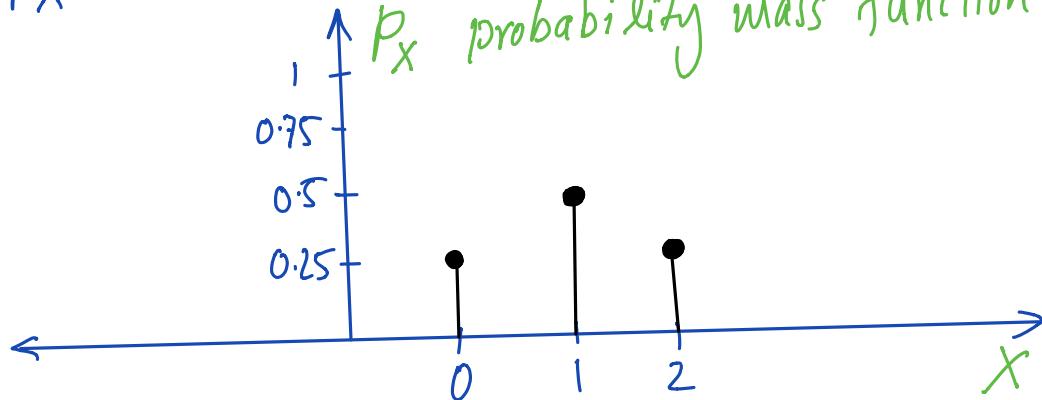
$$P_X(0) = P(X=0) = \frac{1}{4},$$

$$P_X(1) = P(X=1) = \frac{1}{2},$$

$$P_X(2) = P(X=2) = \frac{1}{4},$$

and $P_X(x) = 0$ for all other $x \in \mathbb{R}$.

P_X probability mass function (PMF)



Random Variables

- Y : the number of Tails.

$$P_Y(0) = \frac{1}{4}, P_Y(1) = \frac{1}{2}, P_Y(2) = \frac{1}{4}.$$

- I : 1 if the first toss is Heads and 0 otherwise.

$$P_I(0) = \frac{1}{2}, P_I(1) = \frac{1}{2}.$$

- Homework: Plot PMF's for Y and I .

Conditional probability

- ▶ Example: On a TV game show, hosted by Monty Hall, a contestant chooses one of three closed doors, two of which have a goat behind them and one of which has a car. Monty, who knows where the car is, then opens one of the two remaining doors. The door he opens always has a goat behind it (he never reveals the car!). If he has a choice, then he picks a door at random with equal probabilities. Monty then offers the contestant the option of switching to the other unopened door.
- ▶ If the contestant's goal is to get the car, should she switch doors?

- let W be the event that she wins.
- We want to find the best of the two strategies to win (i.e., the one with higher pr. of winning.)
- Strategies : ① Switch ② No switch.

Conditional probability

- Let C_i be the event that the car is behind door i .
- Assume without loss of generality that she picks the door 1 (else we can relabel the doors)
rename
- Then,
 - ① $P(W) = \underbrace{P(W|C_1) \cdot P(C_1)}_{\text{since switched from the door 1 (given that the car}} + P(W|C_2) \cdot P(C_2) + P(W|C_3) \cdot P(C_3)$
 - ② $P(W) = P(W|C_1) \cdot P(C_1) + P(W|C_2) \cdot P(C_2) + P(W|C_3) \cdot P(C_3)$
- Hence, to switch is the best strategy (from ① & ②).

Lecture 8: Discrete Random Variables - Part II

Satyajit Thakor
IIT Mandi

6 March, 2020

Recall: summary of main ideas

of an experiment

- A sample space S is the set of all possible outcomes.
$$S = \{ s : s \text{ is an outcome} \}$$

- “An event A occurred” if
the outcome of the experiment is in A .

- Probability is
a function from the set of events to the set $[0, 1]$
(satisfying Axioms 1 & 2).
$$P: \{ A \subseteq S : A \text{ is an event} \} \rightarrow [0, 1].$$

- A random variable X is
a function from the sample space S to the set of real
numbers.
$$X: S \rightarrow \mathbb{R}.$$

Recall: summary of main ideas

- ▶ Probability mass function (PMF) p_X of a discrete r.v. X is the function $p_X(x) = P(X=x)$.

$$p_X : \mathbb{R} \rightarrow [0, 1]$$

- ▶ Support of a discrete r.v. X is the set of all values x s.t.

$$p_X(x) > 0.$$

$$\text{support of } X = \{x \in \mathbb{R} : p_X(x) > 0\}.$$

- ▶ **Caution:** For precise definition of the terms refer to the earlier lecture slides.

Discrete random variables

- ▶ Example: Roll two fair 6-sided dice. Let $T = X + Y$ be the total of the two rolls, where X and Y are the individual rolls. What is the PMF of T ?

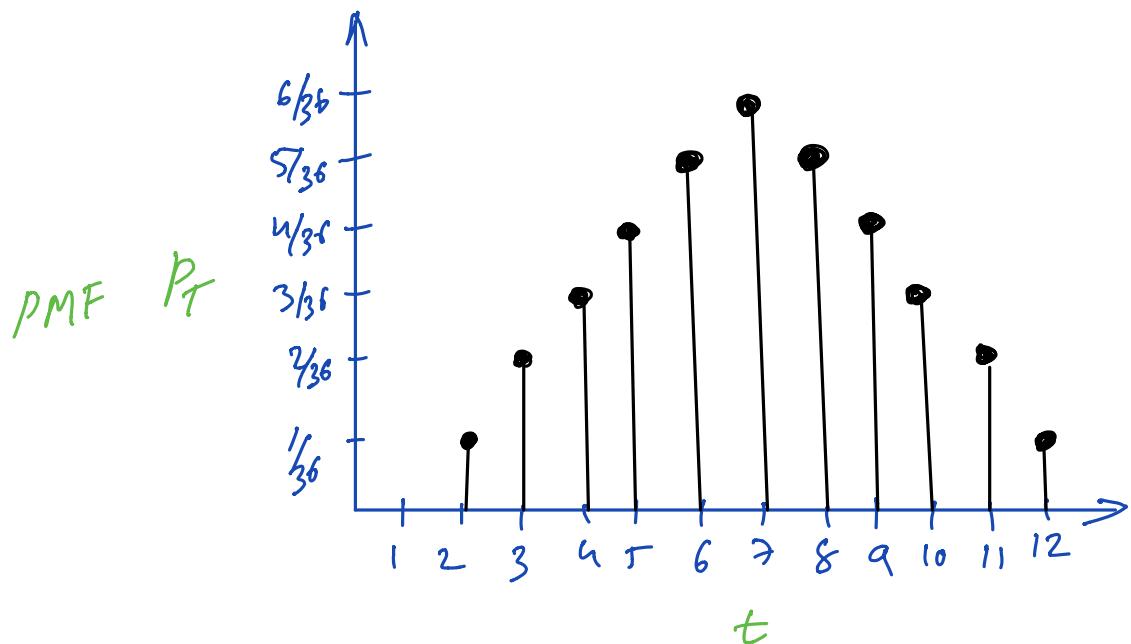
- What values T can take?: support of $T = \{2, 3, \dots, 12\}$.

$$\begin{aligned} - p_T(2) &= P(T=2) = P(\{X=1, Y=1\}) \\ &= P(X=1) \cdot P(Y=1) \\ &= \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} \text{ "U"} \\ - p_T(3) &= P(T=3) = P(\underbrace{\{X=1, Y=2\}}_{\text{disjoint events}} \text{ or } \underbrace{\{X=2, Y=1\}}_{\text{disjoint events}}) \\ &= P(\{X=1, Y=2\}) + P(\{X=2, Y=1\}) \\ &= \frac{2}{36} = p_T(1) \end{aligned}$$

Discrete random variables

- Similarly, $P_T(4) = 3/36 = P_T(10)$
 $P_T(5) = 4/36 = P_T(9)$
 $P_T(6) = 5/36 = P_T(8)$
 $P_T(7) = 6/36$

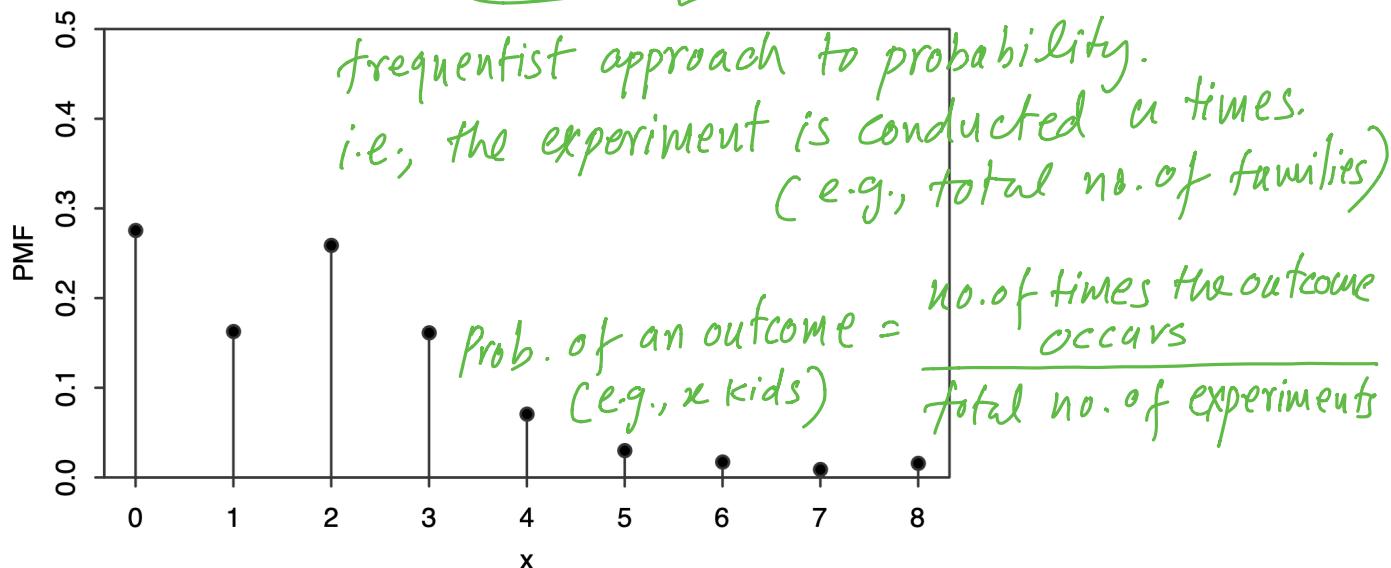
} Homework:
Verify.



Discrete random variables

- ▶ Example of data: Suppose we choose a family in a town at random. Let X be the number of children in the chosen family. Since X can only take on integer values, it is a discrete r.v. The probability that X takes on the value x is proportional to the number of families in the town with x children, i.e.,

$$p_X(x) = P(X = x) = \frac{\text{no. of families with } x \text{ kids}}{\text{total no. of families}}.$$



Discrete random variables

i.e., PMF is valid if ① & ② are satisfied.

- Theorem: Let X be a discrete r.v. with support $\{x_1, x_2, \dots\}$. The PMF p_X of X must satisfy the following two criteria:

1. Nonnegative: $p_X(x) > 0$ if $x = x_j$ for some j , and $p_X(x) = 0$ otherwise;
 2. Sums to 1: $\sum_{i=1}^{\infty} p_X(x_i) = 1$

Proof: ① follows from the defⁿ of support.
 ②: $\sum_{i=1}^{\infty} p_X(x_i) = P\left(\bigcup_{i=1}^{\infty} \{X=x_i\}\right) = P(X=x_1 \text{ or } X=x_2 \text{ or } \dots) = 1$
 by Axiom 2.

- ▶ Conversely, if distinct values $\{x_1, x_2, \dots\}$ are specified and we have a function satisfying the two criteria above, then this function is the PMF of some r.v.

- follows from the defⁿ of discrete r.v.

Discrete random variables

- ▶ Example: Let T be the sum of two fair die rolls. We have already calculated the PMF of T . What is the probability that T is in the interval $[1, 4]$?

$$\begin{aligned} P(T \in [1, 4]) &= P(\{T=2\} \text{ or } \{T=3\} \text{ or } \{T=4\}) \\ &= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} \\ &= \frac{1}{6} \end{aligned}$$

Discrete random variables

- ▶ An r.v. X is said to have the Bernoulli distribution with parameter p if $P(X = 1) = p$ and $P(X = 0) = 1 - p$, where $0 < p < 1$. We write this as $X \sim \text{Bern}(p)$.
- ▶ The symbol \sim is read “is distributed as”.
- ▶ The number p in $\text{Bern}(p)$ is called the parameter of the distribution.
- ▶ Examples: $X \sim \text{Bern}(1/3)$, $Y \sim \text{Bern}(1/8)$
- ▶ The indicator random variable of an event A is the r.v. which equals 1 if A occurs and 0 otherwise.
- ▶ The indicator r.v. of A is denoted I_A or $I(A)$.
- ▶ Note that $I_A \sim \text{Bern}(p)$ with $p = P(A)$.

Discrete random variables

- ▶ An experiment that can result in either a “success” or a “failure” (but not both) is called a Bernoulli trial.
- ▶ A Bernoulli random variable can be thought of as the indicator of success in a Bernoulli trial: it equals 1 if success occurs and 0 if failure occurs in the trial.
- ▶ Suppose that n independent Bernoulli trials are performed, each with the same success probability p . Let X be the number of successes. The distribution of X is called the Binomial distribution with parameters n and p .
- ▶ $X \sim \text{Bin}(n, p)$ denotes that X has the Binomial distribution with parameters n and p , where n is a positive integer and $0 < p < 1$.

Lecture 9: Discrete Random Variables - Part III

Satyajit Thakor
IIT Mandi

10 March, 2020

Discrete random variables

$$(x_i \sim \text{Bern}(p), s=1, f=0)$$

success failure

Ex: Let $n=3, k=1$. A sequence is (s, f, f)
 $\rightarrow \binom{3}{1} = 3$ possible sequences with one s :
 $(s, f, f), (t, s, f), (f, t, s)$

$$P(X_1=s, X_2=f, X_3=f)$$

$$= P(X_1=s) \cdot P(X_2=f) \cdot P(X_3=f) = p \cdot (1-p)^2$$

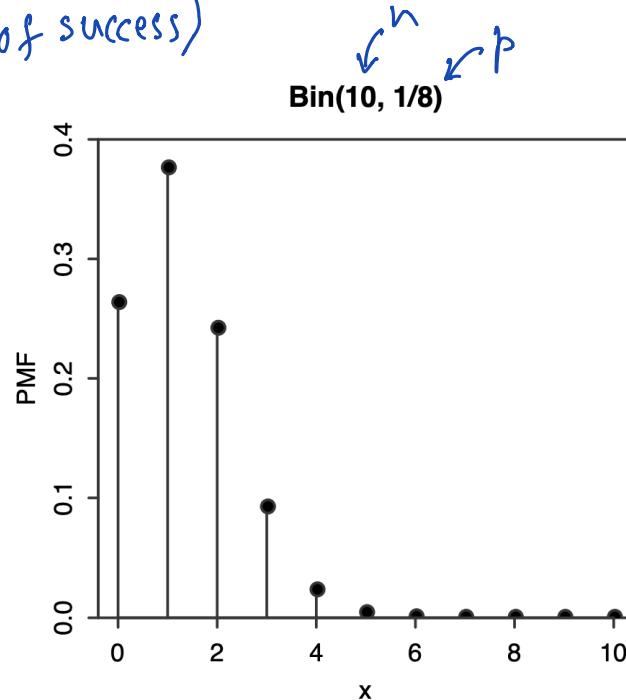
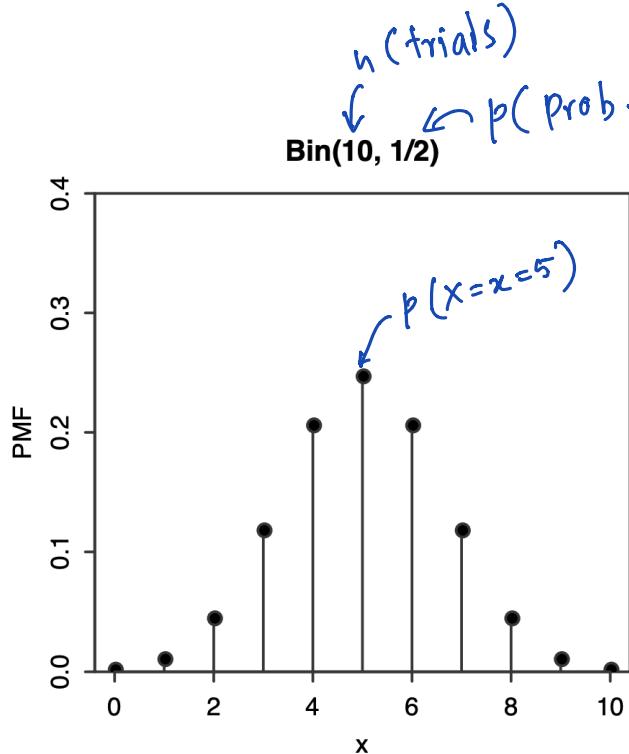
- Theorem: If $X \sim \text{Bin}(n, p)$, then the PMF of X is

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, \dots, n$ and $p_X(k) = P(X = k) = 0$ otherwise.

- Prob. that $X=k$ is the prob. that in n Bernoulli trials we have k successes (and $n-k$ failures)
- There are $\binom{n}{k}$ such ways (sequences/strings) s.t. k successes can occur in n Bernoulli trials.
- Prob. of occurrence of such a sequence is $p^k \cdot (1-p)^{n-k}$.
- Hence, $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k=0, 1, \dots, n$ and 0 otherwise.

Discrete random variables



- ▶ Note that the PMF of the $\text{Bin}(10, 1/2)$ distribution is symmetric about 5, but when the success probability is not $1/2$, the PMF is skewed.

Discrete random variables

i.e., $0 < p < 1$
"open interval"
 $(0, 1) \triangleq \{x \in \mathbb{R} : 0 < x < 1\}$.

- ▶ Consider a sequence of independent Bernoulli trials, each with the same success probability $p \in (0, 1)$, with trials performed until a success occurs. Let X be the number of failures before the first successful trial. Then X has the Geometric distribution with parameter p ; denoted $X \sim \text{Geom}(p)$.
- ▶ Theorem: If $X \sim \text{Geom}(p)$, then the PMF of X is

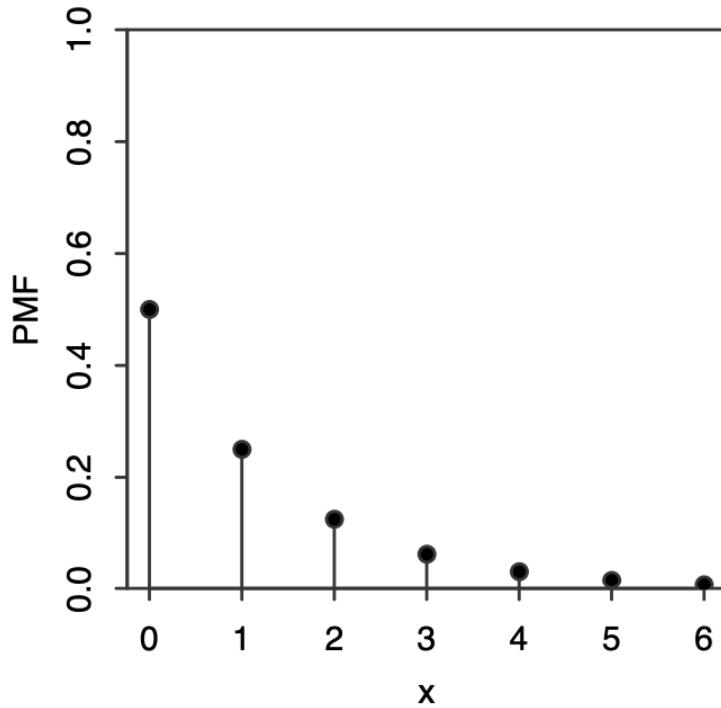
$$P(X = k) = (1 - p)^k p, \quad \text{for } k = 0, 1, 2, \dots$$

- ▶ Note that, the support of a geometric r.v. has infinite cardinality.

- For $0 < p < 1$, the value $(1-p)^k \cdot p$ is always a positive number in the set $(0, 1)$.

Discrete random variables

- Geom($1/2$):



- Theorem: $\text{Geom}(p)$ is a valid PMF.

- We need to verify whether "non-negative" and "sums to 1" properties are satisfied.

Discrete random variables

- Note that $P_X(k) \geq 0$. Now, we only need to verify whether

$$\sum_{k \in \mathbb{R}} P_X(k) = 1.$$

$$\begin{aligned}\sum_{k \in \mathbb{R}} P_X(k) &= \sum_{k=0}^{\infty} P_X(k) \\&= \sum_{k=0}^{\infty} p (1-p)^k, \quad k=0, 1, 2, \dots \\&= p \underbrace{\sum_{k=0}^{\infty} (1-p)^k}_{\text{geometric series}} \\&= p \cdot \frac{1}{(1-(1-p))} \\&= 1.\end{aligned}$$

Geometric series

$$S_n = 1 + y + y^2 + \dots + y^n, \quad y \in (0, 1)$$

$$y \cdot S_n = y + y^2 + \dots + y^n + y^{n+1}$$

$$\Rightarrow S_n (1-y) = 1 - y^{n+1}$$

$$\Rightarrow S_n = \frac{1 - y^{n+1}}{1-y}$$

As $n \rightarrow \infty$ we have $y^{n+1} \rightarrow 0$
since $y \in (0, 1)$.

$$\Rightarrow S_{\infty} = \frac{1}{1-y} = \sum_{k=0}^{\infty} y^k.$$

Discrete random variables

- ▶ Let C be a finite, non-empty set of numbers. Choose one of these numbers uniformly at random (i.e., all values in C are equally likely). Call the chosen number X . Then X is said to have the Discrete Uniform distribution with parameter C ; we denote this by $X \sim \text{DUnif}(C)$.
- ▶ The PMF of $X \sim \text{DUnif}(C)$ is

$$P(X = x) = \frac{1}{|C|}$$

- ▶ Note that, for any $A \subseteq C$,

$$P(X \in A) = \frac{|A|}{|C|}.$$

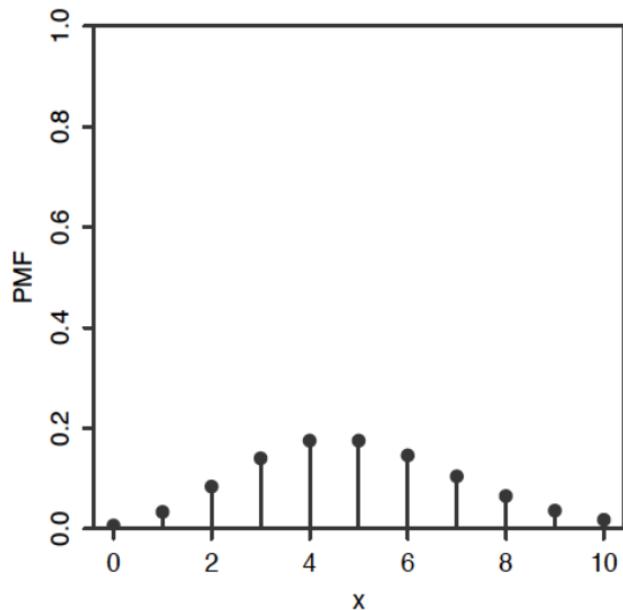
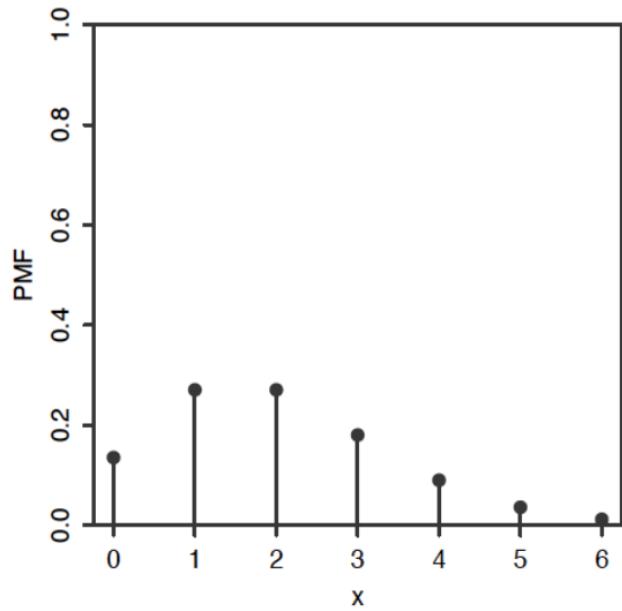
- ▶ Example: Imagine an n -sided fair dice.

Discrete random variables

- ▶ Many random phenomena observed can be modelled as Poisson distribution, e.g., the number of customers arriving at a shopping centre within a fixed interval.
- ▶ An r.v. X has the Poisson distribution with parameter λ , where $\lambda > 0$, if the PMF of X is
*any real no.
greater than 0.*
$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$
- ▶ Written as $X \sim \text{Pois}(\lambda)$.

Discrete random variables

- ▶ Example: Pois(2) and Pois(5).



- ▶ Theorem: The Poisson distribution is a valid distribution.

Proof: First note that $p(X=k) = \frac{e^{-\lambda} \lambda^k}{k!}$ is non-negative for all $k \in \mathbb{R}$.

Discrete random variables

- Now,

$$\sum_{k \in \mathbb{N}} P_X(k) = \sum_{k=0}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!}, \quad k=0,1,2,\dots$$

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

Taylor series (expansion)
of e^x .

Taylor series of a differentiable function $f(x)$ centered at a :

$$f(x) = f(a) + \frac{f'(a)}{1!} (x-a) + \frac{f''(a)}{2!} (x-a)^2 + \dots$$

$$= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n \quad \text{where } f^{(n)}(a) = \left. \frac{d^n f(x)}{dx^n} \right|_{x=a}$$

Let $f(x) = e^x$, $a=0$. Then,

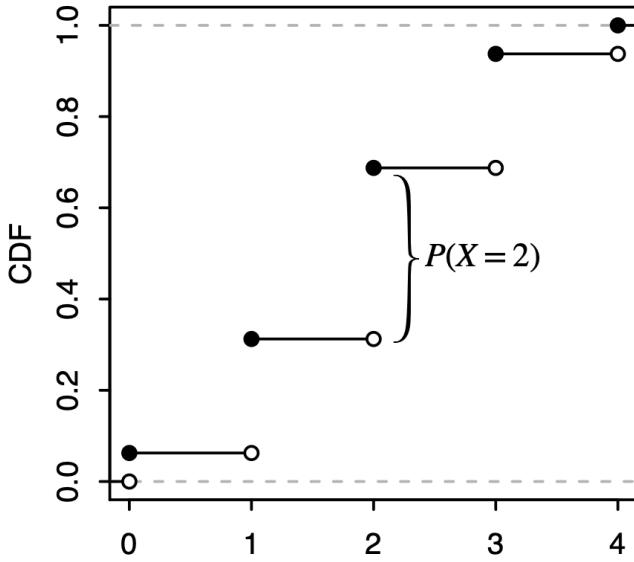
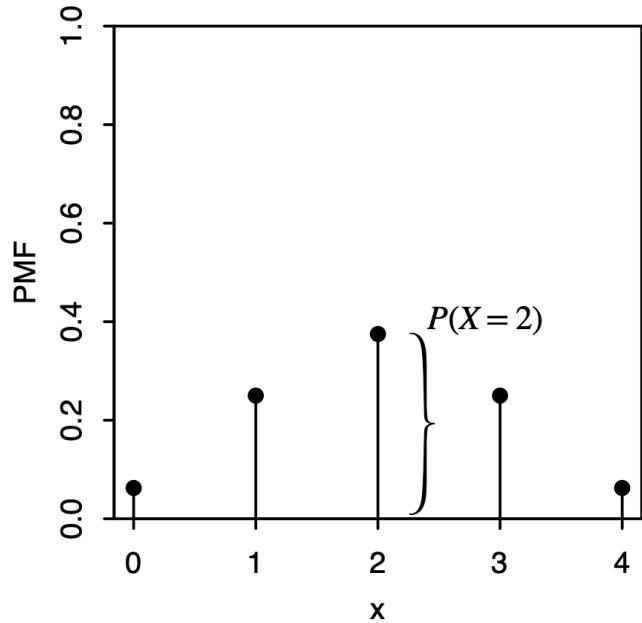
$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Discrete random variables

- ▶ Another function that describes the distribution of an r.v. is the cumulative distribution function (CDF).
- ▶ Unlike the PMF, which only discrete r.v.s have, the CDF is defined for all r.v.s.
- ▶ The cumulative distribution function (CDF) of an r.v. X is the function $F_X(x) = P(X \leq x)$.

Discrete random variables

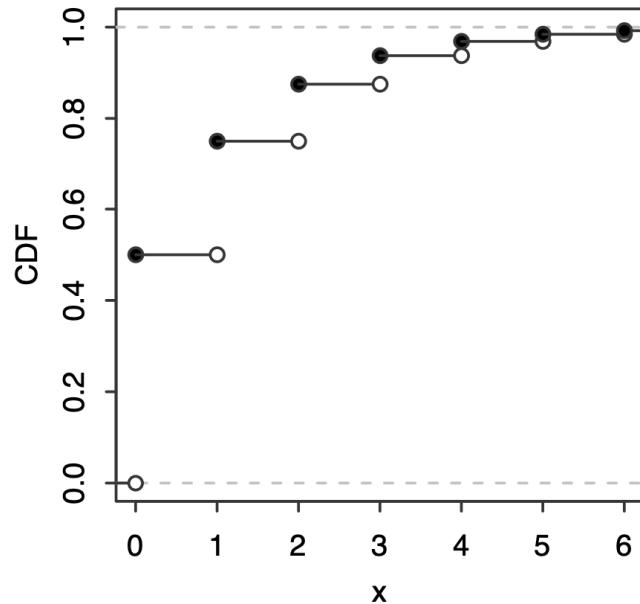
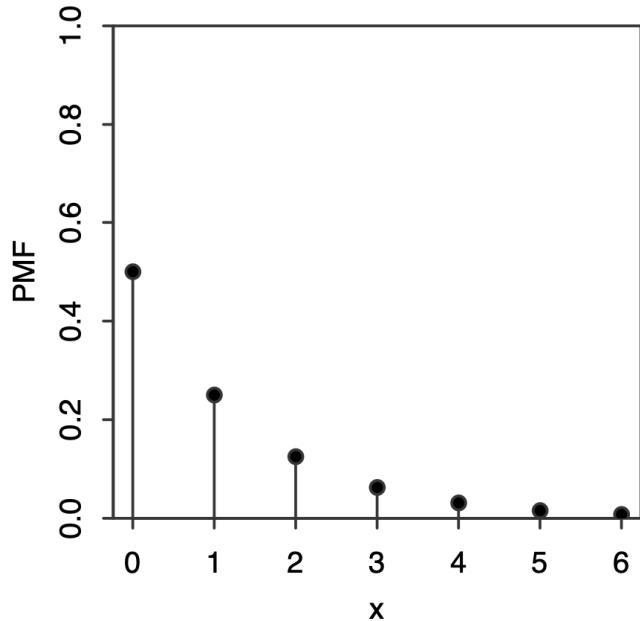
- ▶ Example: $\text{Bin}(4, 1/2)$



- ▶ From PMF to CDF $F_X(1.5) = P(X \leq 1.5) = p(X=0) + p(X=1) + p(1 < X \leq 1.5) = (\frac{1}{2})^4 + 4(\frac{1}{2})^4 + 0 = 5/16$.
- ▶ From CDF to PMF $P_X(1) = P(X=1) = P(X \leq 1) - \underbrace{P(X < 1)}_{= P(X=0)} = \frac{5}{16} - \frac{1}{16} = \frac{4}{16} = \frac{P(X=0) + P(0 < X < 1)}{16} = \frac{1}{16} + 0$

Discrete random variables

- ▶ Example: $\text{Geom}(1/2)$



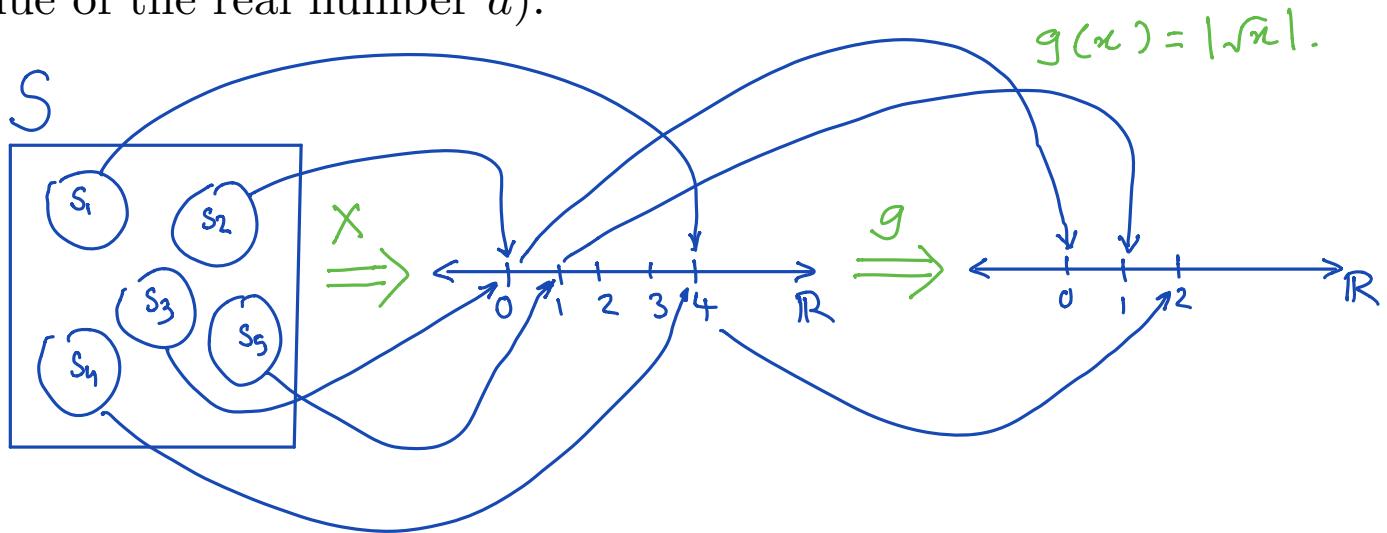
Lecture 10:
Discrete Random Variables - Part IV
&
Mini Quiz - I

Satyajit Thakor
IIT Mandi

13 March, 2020

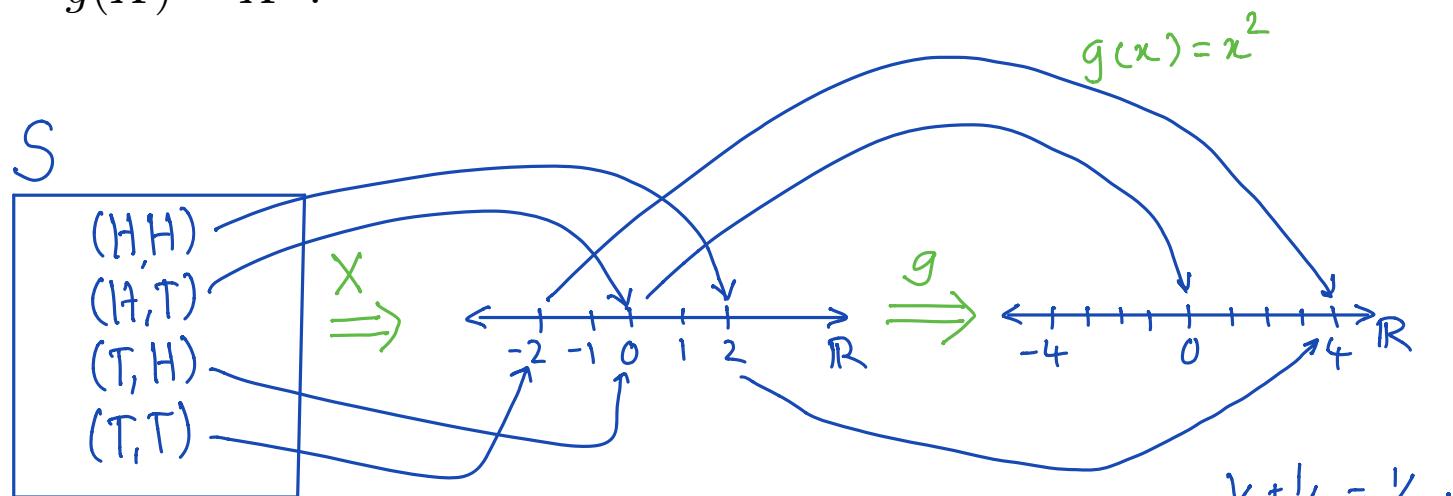
Discrete random variables

- ▶ If X is a random variable, then X^2 , e^X , and $\sin(X)$ are also random variables.
- ▶ Definition (Function of an r.v.): For an experiment with sample space S , an r.v. X , and a function $g : \mathbb{R} \rightarrow \mathbb{R}$, $g(X)$ is the r.v. that maps s to $g(X(s))$ for all $s \in S$.
- ▶ Example: $g(x) = |\sqrt{x}|$ (here the notation $|a|$ means the absolute value of the real number a).



Discrete random variables

- Example: Two fair coins are tossed. Let X denote the following r.v.: whenever a head occurs Rs 1 is gained and whenever a tail occurs the same amount is lost. What is the PMF for $Y = g(X) = X^2$?



$$p_Y(0) = P(g(X)=0) = P(X=0) = P((H,T) \text{ or } (T,H)) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$
$$p_Y(4) = P(g(X)=4) = P(\{X=-2\} \cup \{X=2\}) = P((T,T) \text{ or } (H,H)) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Discrete random variables

- Let X be an r.v. with PMF $p_X(x)$ and $Y = g(X)$. If g is a one-to-one function what is the PMF of Y ?

Let $p_i = P(X=x_i)$ and $y_i = g(x_i)$. Then, $\underbrace{P(Y=y_i)}_{\because g \text{ is a one-to-one function}} = \underbrace{P(X=x_i)}_{= p_i} = p_i$

- Let X be an r.v. with PMF $p_X(x)$ and $Y = g(X)$. If g is a function (not necessarily one-to-one) what is the PMF of Y ?

$$\begin{aligned} p_Y(y) &= P(Y=y) = P(g(X)=y) \\ &= P\left(\bigcup_{x: g(x)=y} \{X=x\}\right) \\ &= \sum_{x: g(x)=y} P(X=x) \quad \text{union for all } x \text{ s.t. } g(x)=y. \\ &= \sum_{x: g(x)=y} p_X(x). \quad \text{sum for all } x \text{ s.t. } g(x)=y. \end{aligned}$$

Discrete random variables

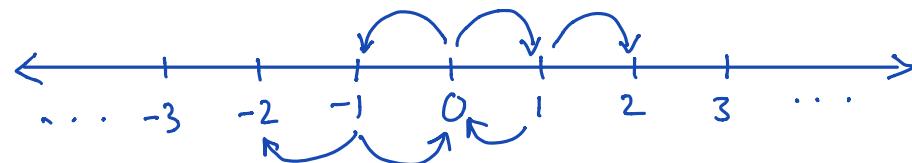
- A particle moves n steps on a number line. The particle starts at 0, and at each step it moves 1 unit to the right or to the left, with equal probabilities. Assume all steps are independent.

(a) Let X be the number of steps taken to the right. Find the PMF of X .

(b) Let Y denote the location of the particle after n steps.

Write Y in terms of X .

(c) Find the PMF of Y .



- The support of X is $\{0, 1, \dots, n\}$.

- What is the PMF of X ? : $\text{Bin}(n, 1/2)$

- Consider Bernoulli trial at each step (n Bernoulli trials)
- Consider moving to the right as success and left as failure.

Discrete random variables

- $P(X=i)$ is equivalent to prob. of i successes in n Bernoulli trials.
- Hence $X \sim \text{Bin}(n, 1/2)$.
- When $X=i$, what is the position of the particle?
moved right i times and moved left $n-i$ times.
 \Rightarrow Final position = $i - (n-i) = 2i - n$ for $i \in \{0, 1, 2, \dots, n\}$
- Let Y be the r.v. that mean the final position of the particle.
Then $Y = 2X - n$, support of $Y = \{-n, \dots, 0, \dots, n\}$.
 $P(Y=j) = P(2X-n=j) = P\left(X=\frac{n+j}{2}\right) = \binom{n}{\frac{n+j}{2}} \left(\frac{1}{2}\right)^n$ for $j \in \{-n, \dots, 0, \dots, n\}$ and 0 otherwise.

Discrete random variables

- Let D be the particle's distance from the origin after n steps.
Find the PMF of D .

Note that, $D = g(Y) = |Y|.$

Support of $D = \{0, 1, \dots, n\}.$

$$P_D(0) = P(D=0) = P(Y=0) = \binom{n}{n/2} \left(\frac{1}{2}\right)^n.$$

$$P_D(k) = P(D=k) = P(\{Y=-k\} \text{ or } \{Y=+k\})$$

$$= P(Y=-k) + P(Y=+k)$$

$$= \left[\binom{n}{\frac{n-k}{2}} + \binom{n}{\frac{n+k}{2}} \right] \left(\frac{1}{2}\right)^n$$

$$\text{for } k \in \{1, 2, \dots, n\}.$$

Mini-Quiz (9 points, 17 mins)

- ▶ If $P(E) = 0.1$ and $P(F) = 0.2$, then find the best lower and upper bounds for $P(E \cup F)$, i.e., find biggest l and smallest u such that $l \leq P(E \cup F) \leq u$. [1 point]
- ▶ Define partition of a set S . [1.5 points]
- ▶ Bag A contains 3 red balls and 7 blue balls. Bag B contains 8 red and 4 blue balls. Bag C contains 5 red and 11 blue balls. A bag is chosen at random, and then a ball is chosen at random from that bag. Calculate the probabilities:
 - (a) A red ball is chosen. [1 point]
 - (b) A red ball from bag B is chosen. [1 point]
 - (c) If it is known that a red ball is chosen, what is the probability that it comes from bag A? [1.5 points]
- ▶ A fair coin is tossed three times. A player wins Rs. 1 if the first toss is a head, but loses Rs. 1 if the first toss is a tail. Similarly, the player wins Rs. 2 if the second toss is a head, but loses Rs. 2 if the second toss is a tail, and wins or loses Rs. 3 according to the result of the third toss. Let the random variable X be the total winnings after the three tosses. Find its PMF [2 points] and plot it [1 point].

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

① $0.2 \leq \overbrace{P(EUF)} \leq 0.3$

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

- ② The subsets A_1, \dots, A_n of S forms a partition of S
 if $\underbrace{A_i \cap A_j = \emptyset}_{i.e., A_1, \dots, A_n}$ for all $i \neq j$ and $\bigcup_{i=1}^n A_i = S$.

- ③ Let R be the event that a red ball is chosen.
 Let B_i be the event that bag i is chosen.

(a) $P(R) = \sum_{i \in \{A, B, C\}} P(R|B_i) \cdot P(B_i)$
 $= \frac{1}{3} \cdot \frac{3}{10} + \frac{1}{3} \cdot \frac{8}{12} + \frac{1}{3} \cdot \frac{5}{16} \approx 0.42$

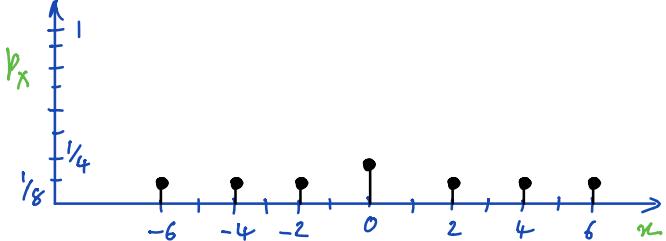
(b) $P(R \cap B_B) = P(B_B) \cdot P(R|B_B)$
 $= \frac{1}{3} \cdot \frac{8}{12} = \frac{2}{9}$

(c) $P(B_A | R) = \frac{P(B_A \cap R)}{P(R)} = \frac{P(B_A) \cdot P(R|B_A)}{P(R)} = \frac{\frac{1}{3} \cdot \frac{3}{10}}{0.42} \approx 0.23$

- ④ outcome amount won (X) $P_X(-6) = P_X(-4) = P_X(-2) = P(X=2) = P(X=4)$

H H H	6
A H T	0
H T H	2
T H H	4
H T T	-4
T H T	-2
T T H	0
T T T	-6

$$P_X(0) = P(\text{HHT or TTH}) = \frac{1}{4}.$$



Lecture 11: Discrete Random Variables - Part V

Satyajit Thakor
IIT Mandi

23 March, 2020

Functions of r.v.s

- ▶ Definition (function of 2 r.v.s): Given an experiment with sample space S , if X and Y are r.v.s that map $s \in S$ to $X(s)$ and $Y(s)$ respectively, then $g(X, Y)$ is the r.v. that maps s to $g(X(s), Y(s))$.
- ▶ Two fair dice are rolled. X is the number shown on 1st die and Y is the number shown on 2nd die. Find the PMF of $Z = \max(X, Y)$.

$$P(\max(X, Y) = 1) = P(\{X=1, Y=1\}) = 1/36.$$
$$P(\max(X, Y) = 2) = P(\{X=1, Y=2\} \cup \{X=2, Y=1\} \cup \{X=2, Y=2\}) = 3/36.$$
$$P(\max(X, Y) = 3) = P(\{X=1, Y=3\} \cup \{X=2, Y=3\} \cup \{X=3, Y=3\} \cup \{X=3, Y=2\} \cup \{X=3, Y=1\}) = 5/36.$$

Similarly, verify that (Homework):

$$P(\max(X, Y) = 4) = 7/36, P(\max(X, Y) = 5) = 9/36, P(\max(X, Y) = 6) = 11/36.$$

Joint distributions

- ▶ Recall: The distribution of a discrete random variable can be described by its PMF and also by CDF.
- ▶ The joint distribution of two (or more) discrete r.v.s can be described by their joint PMF and also by joint CDF.
- ▶ The joint PMF of discrete r.v.s X and Y is the function $p_{X,Y} : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$ given by

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

Example:

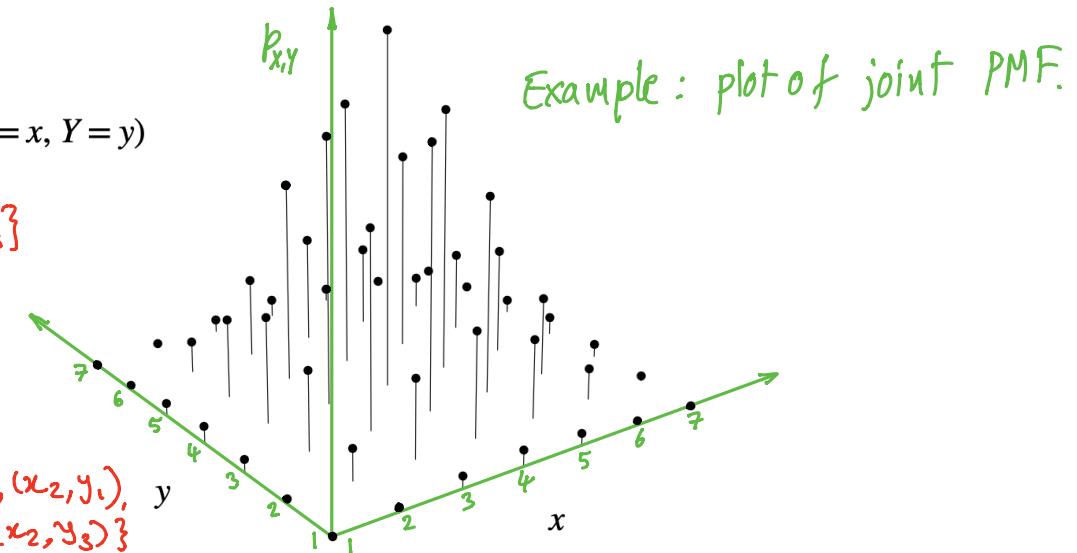
alphabet of $X = \{x_1, x_2\}$

alphabet of $Y = \{y_1, y_2, y_3\}$

What is the alphabet of the joint r.v. (X, Y) ?

$\{x_1, x_2\} \times \{y_1, y_2, y_3\}$

$= \{(x_1, y_1), (x_1, y_2), (x_1, y_3), (x_2, y_1), (x_2, y_2), (x_2, y_3)\}$



Joint distributions

- ▶ Example (Two Bernoulli r.v.s X and Y):

The joint PMF can be completely specified by 4 values:

These are two alternative ways to express joint PMF.

$$\left\{ \begin{array}{l} P(X=1, Y=1), P(X=0, Y=1), P(X=1, Y=0), P(X=0, Y=0) \\ P(X=1, Y=1), P(X=0, Y=1), P(X=1, Y=0), P(X=0, Y=0) \end{array} \right.$$

Let $= \frac{5}{100}$ $= \frac{3}{100}$ $= \frac{20}{100}$ $= \frac{72}{100}$.

Then the joint PMF can be expressed in the table form:

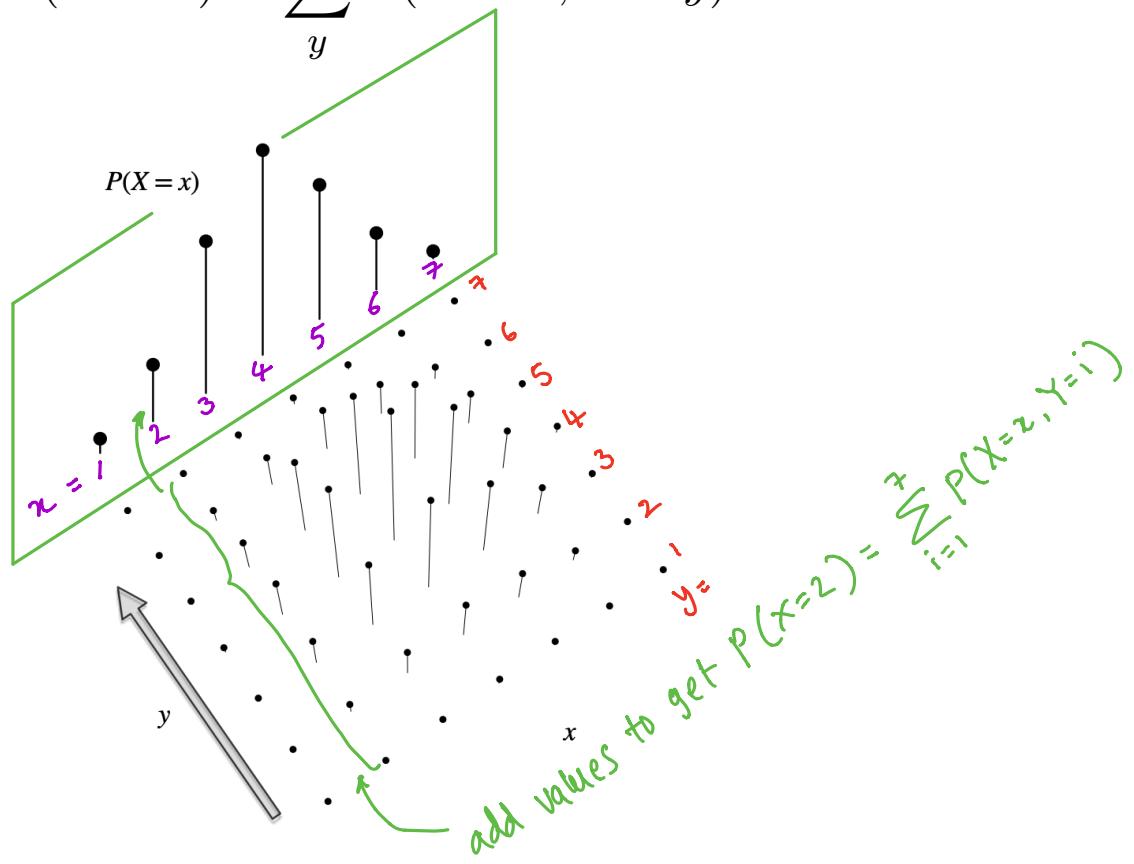
		$Y=1$	$Y=0$
$X=1$	1	0.05	0.2
	0	0.03	0.72

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

Joint distributions

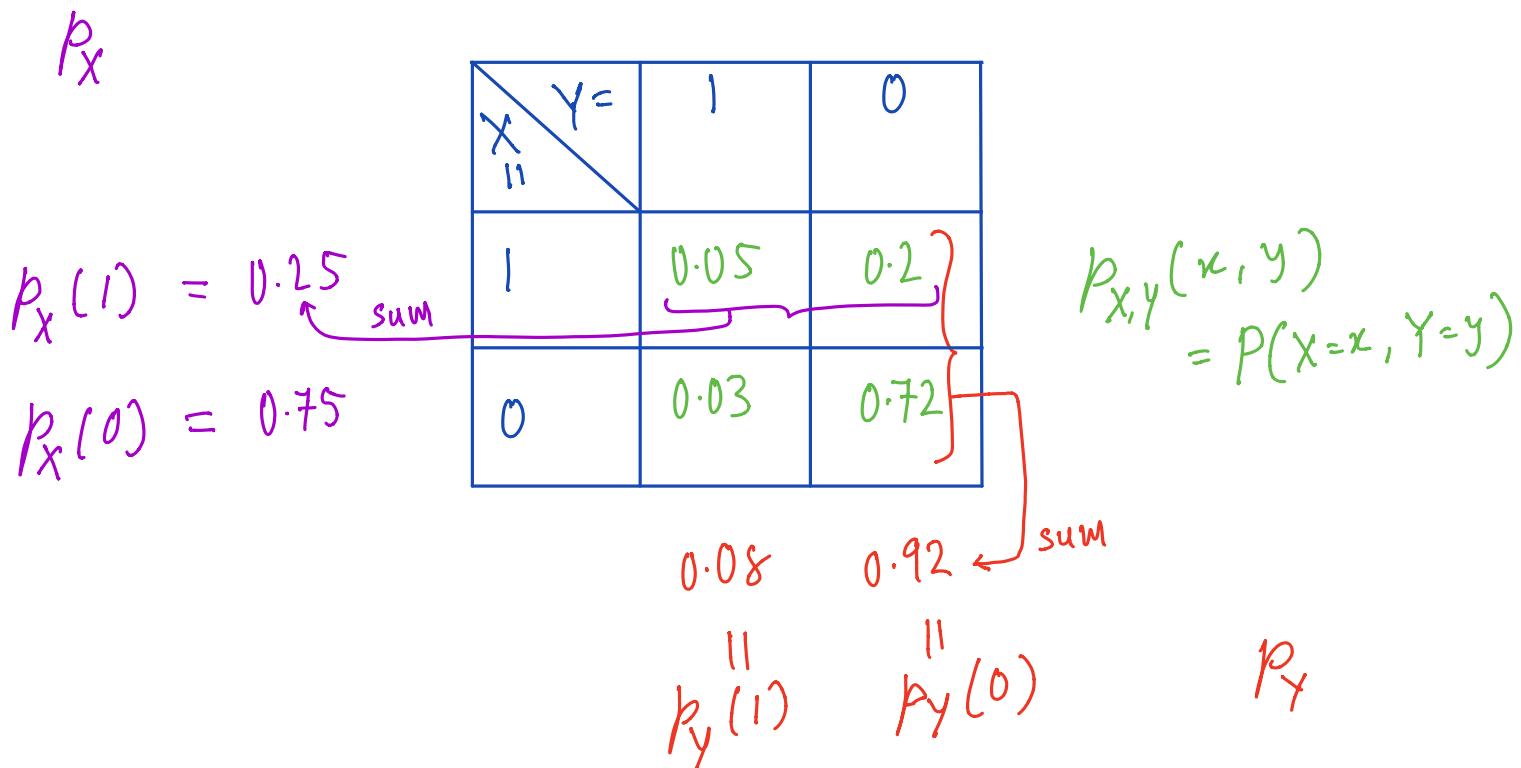
- For discrete r.v.s X and Y , the marginal PMF of X is

$$P(X = x) = \sum_y P(X = x, Y = y).$$



Joint distributions

- ▶ Example (Two Bernoulli r.v.s X and Y):



Joint distributions

- The joint CDF of r.v.s X and Y is the function $F_{X,Y} : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$ given by

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

- Example (Two Bernoulli r.v.s X and Y):

$$F_{X,Y}(0, 0) = P(X \leq 0, Y \leq 0) = 0.72$$

$$F_{X,Y}(0, 1) = \underbrace{P(X \leq 0, Y \leq 1)}_{P(\{X=0, Y=0\} \cup \{X=0, Y=1\})} = 0.75$$

$$F_{X,Y}(1, 0) = \underbrace{P(X \leq 1, Y \leq 0)}_{P(\{X=0, Y=0\} \cup \{X=1, Y=0\})} = 0.94$$

Similarly, $F_{X,Y}(1, 1) = 1$.
Note that, $F_{X,Y}(x, y) = 0$ for all $x < 0$ or $y < 0$.
 $F_{X,Y}(x, y) = 1$ for all $x \geq 1$ and $y \geq 1$.

Independent r.v.s

- ▶ Recall (independent events): The events A and B are independent if $P(A \cap B) = P(A)P(B)$.
- ▶ Random variables X and Y are said to be **independent** if

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y),$$

for all $x, y \in \mathbb{R}$.

- ▶ For discrete r.v.s, this is equivalent to

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

Independent r.v.s

- Random variables X_1, \dots, X_n are independent if

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = P(X_1 \leq x_1) \dots P(X_n \leq x_n),$$

for all $x_1, \dots, x_n \in \mathbb{R}$.

- Recall: For n events to be independent, 2^n (one equality for each subset $J \subseteq \{1, 2, \dots, n\}$) equalities must be satisfied.
- But for n r.v.s to be independent only one equality must be satisfied. In fact, this one equality implies all other equalities for subsets of $\{1, 2, \dots, n\}$.



- This follows from certain properties of CDF.
(Advance : details omitted)

Lecture 12: Discrete Random Variables - Part VI

Satyajit Thakor
IIT Mandi

25 March, 2020

Independent r.v.s

- Example (Two Bernoulli r.v.s X and Y): Are they independent?

$p_X :$

$$p_X(1) = 0.25$$

$$p_X(0) = 0.75$$

$\begin{matrix} X \\ \backslash \\ 1 \end{matrix}$	$Y=1$	$Y=0$
$Y=1$	0.05	0.2
$Y=0$	0.03	0.72

$$\begin{aligned} p_{X,Y}(x,y) \\ = P(X=x, Y=y) \end{aligned}$$

$$p_Y : p_Y(1) = 0.08, p_Y(0) = 0.92$$

- Note that, $P(X=1, Y=1) = 0.5$
 $P(X=1) \cdot P(Y=1) = 0.25 \cdot 0.08 = 0.02$
 $\Rightarrow P(X=1, Y=1) \neq P(X=1) \cdot P(Y=1)$
 $\Rightarrow X$ and Y are not independent.

Independent r.v.s

- In a roll of two fair dice, Let X be the number on the first die and Y be the number on the second die. Consider functions $g(X, Y) = X + Y$ and $h(X, Y) = X - Y$ and denoted them as r.v.s G and H .
- Find p_G, p_H (directly).
 - Find $p_{G,H}$ and write in the table form. Then find marginal PMFs p_G, p_H from the joint PMF and verify your solution with the solution of (a).
 - Are $G = g(X, Y)$ and $H = h(X, Y)$ independent?

(a): - To find p_G and p_H , we need to know the support of G, H and no. of ways each element can occur.

- The support of $X+Y$ is $\{2, 3, \dots, 12\}$
- The support of $X-Y$ is $\{-5, -4, \dots, 4, 5\}$

Independent r.v.s

$g = \text{Value}(g)$: (x, y) :	no. of ways the value can occur:	$P_G(g)$:
$X+Y = 2$	$(1,1)$	1 $\frac{1}{36}$
3	$(1,2), (2,1)$	2 $\frac{2}{36}$
4	$(1,3), (3,1), (2,2)$	3 $\frac{3}{36}$
5	$(1,4), (4,1), (2,3), (3,2)$	4 $\frac{4}{36}$
6	$(1,5), (5,1), (2,4), (4,2), (3,3)$	5 $\frac{5}{36}$
7	$(1,6), (6,1), (2,5), (5,2), (3,4), (4,3)$	6 $\frac{6}{36}$
8	$(2,6), (6,2), (3,5), (5,3), (4,4)$	5 $\frac{5}{36}$
9	$(3,6), (6,3), (4,5), (5,4)$	4 $\frac{4}{36}$
10	$(4,6), (6,4), (5,5)$	3 $\frac{3}{36}$
11	$(5,6), (6,5)$	2 $\frac{2}{36}$
12	$(6,6)$	1 $\frac{1}{36}$

Independent r.v.s

$H = X - Y$	Value (h): (x, y) :	no. of ways the value can occur:	$p_H(h)$:
-5	$(1, 6)$	1	$\frac{1}{36}$
-4	$(1, 5), (2, 6)$	2	$\frac{2}{36}$
-3	$(1, 4), (2, 5), (3, 6)$	3	$\frac{3}{36}$
-2	$(1, 3), (2, 4), (3, 5), (4, 6)$	4	$\frac{4}{36}$
-1	$(1, 2), (2, 3), (3, 4), (4, 5), (5, 6)$	5	$\frac{5}{36}$
0	$(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)$	6	$\frac{6}{36}$
1	$(2, 1), (3, 2), (4, 3), (5, 4), (6, 5)$	5	$\frac{5}{36}$
2	$(3, 1), (4, 2), (5, 3)$	4	$\frac{4}{36}$
3	$(4, 1), (5, 2), (6, 3)$	3	$\frac{3}{36}$
4	$(5, 1), (6, 2)$	2	$\frac{2}{36}$
5	$(6, 1)$	1	$\frac{1}{36}$

Independent r.v.s

(b):

- Note that $g=2$ if $(x,y) = (1,1)$.
- But if $(x,y) = (1,1)$ then $h=0$.
- Hence $P(G=2, H=0) = P((X,Y) = (1,1)) = \frac{1}{36}$.
- Note that $P(G=2, H=i) = 0$ for all $i \neq 0$.
- with this reasoning we have the first column ($g=2$).
- Similarly, we find $P_{GH}(g,h)$ for other values of g and h .

Homework: Verify all the values in the table.
 (Use the reasoning similar to one in red text)

G	1	2	3	4	5	6	7	8	9	10	11	12	P_H
$h \backslash g$	0	0	0	0	0	$\frac{1}{36}$	0	0	0	0	0	0	$\frac{1}{36}$
-5	0	0	0	0	0	$\frac{1}{36}$	0	0	0	0	0	0	$\frac{1}{36}$
-4	0	0	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	0	0	0	$\frac{1}{36}$
-3	0	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	0	0	$\frac{1}{36}$
-2	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	0	$\frac{1}{36}$
-1	0	$\frac{1}{36}$	0	$\frac{1}{36}$	$\frac{1}{36}$								
0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
1	0	$\frac{1}{36}$	0	$\frac{1}{36}$	$\frac{1}{36}$								
2	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$
3	0	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	0	0	$\frac{1}{36}$
4	0	0	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	0	0	0	$\frac{1}{36}$
5	0	0	0	0	0	$\frac{1}{36}$	0	0	0	0	0	0	$\frac{1}{36}$
P_G	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{7}{36}$	$\frac{8}{36}$	$\frac{9}{36}$	$\frac{10}{36}$	$\frac{11}{36}$	$\frac{12}{36}$	

$P_{G,H}$

Independent r.v.s

(c) :

- By defⁿ: If $P(X+Y=x+y, X-Y=x-y) \neq P(X+Y=x+y) \cdot P(X-Y=x-y)$ for some (x, y) then $X+Y$ and $X-Y$ are not independent.
- Intuitively, we can see dependence between $X+Y$ and $X-Y$: For example, if $\underbrace{X+Y=12}_{\text{if } (x,y)=(6,6)}$ then $X-Y$ must be 0.
- Now, note that : $P(X+Y=12, X-Y=1) = 0$
 $P(\underbrace{X+Y=12}_{\text{if } (x,y)=(6,6)}) \cdot P(\underbrace{X-Y=1}_{(x,y)=(2,1) \text{ or } (3,2) \text{ or } (4,3) \text{ or } (5,4) \text{ or } (6,5)}) = \frac{1}{36} \cdot \frac{5}{36}$
- Hence, X and Y are not independent.

Example: joint and marginal distributions

- ▶ Recall: The marginal PMF of X is
 $p_X(x) = P(X = x) = \sum_y P(X = x, Y = y).$
- ▶ Similarly, the marginal CDF of X is
 $F_X(x) = P(X \leq x) = \sum_y P(X \leq x, Y = y).$
- ▶ Example (Two Bernoulli r.v.s X and Y): Find $F_X(0)$.

$X \backslash Y$	1	0
1	0.05	0.2
0	0.03	0.72

$$\begin{aligned} F_X(0) &= P(X \leq 0) \\ &= P(X = 0) \\ &= P(X = 0, Y = 0) + P(X = 0, Y = 1) \\ &= 0.72 + 0.03 \\ &= 0.75. \end{aligned}$$

Lecture 13: Discrete Random Variables - Part VII

Satyajit Thakor
IIT Mandi

27 March, 2020

Conditional distribution

- ▶ Recall: Conditional probability is defined for events.
- ▶ Now we see its generalization: conditional distributions for r.v.s.
- ▶ For discrete r.v.s X and Y , the conditional PMF of Y given $X = x$ is

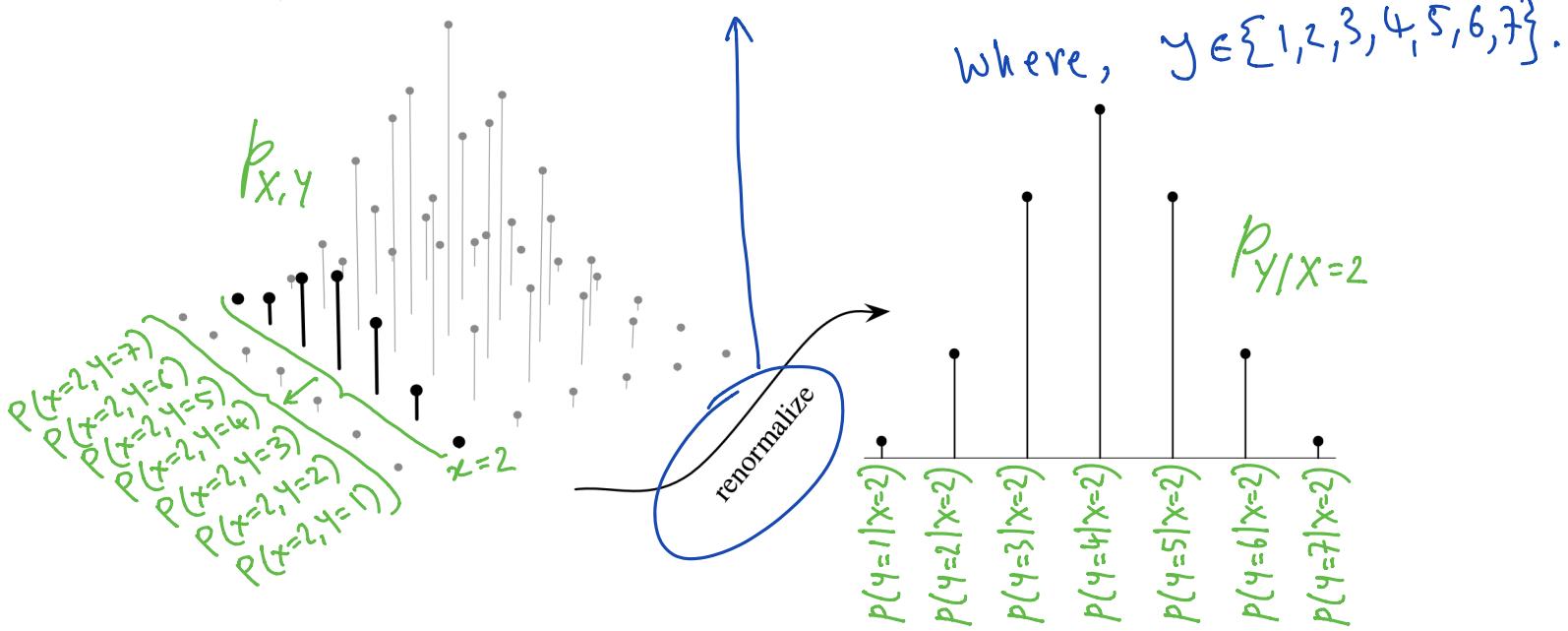
$$p_{Y|X=x}(y) = P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}, \quad P(X = x) > 0.$$

- ▶ The conditional PMF is a function of y for fixed x .
- ▶ Similar to conditional PDF, the conditional CDF of Y given $X = x$ is defined as

$$F_{Y|X=x}(y) = P(Y \leq y|X = x) = \frac{P(X = x, Y \leq y)}{P(X = x)}, \quad P(X = x) > 0.$$

Conditional distribution

i.e., dividing $p(X=2, Y=y)$ by $p(X=2)$: $p(Y=y | X=2) = \frac{p(X=2, Y=y)}{p(X=2)}$



- By renormalization, note that $\sum_y p(Y=y | X=x) = \frac{\sum_y p(X=x, Y=y)}{p(X=x)} = 1.$ (i.e., adds up to 1)
- Hence, conditional PMF too is a PMF.

Conditional distribution

- Two pens are selected at random from a box that contains 3 blue pens, 2 red pens, and 3 green pens. If X is the number of blue pens selected and Y is the number of red pens selected, find
 - (a) the joint PMF $p_{X,Y}$ and the marginal PMFs p_X, p_Y ,
 - (b) $P((X, Y) \in A)$, where A is the region $\{(x, y) : x + y \leq 1\}$,
 - (c) the conditional PMF of X , given that $Y = 1$,
 - (c) the conditional CDF $F_{X|Y=1}(1)$.

(a)- Note that, support of $X = \{0, 1, 2\} = \text{support of } Y$.
- Hence, (x, y) can take values: $(0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), (2,2)$.
- But note that, $\underbrace{P_{X,Y}((1,2)) = P_{X,Y}((2,1)) = P_{X,Y}((2,2)) = 0}_{\because \text{only 2 pens are selected (total)}}$.
- The cardinality of the sample space is $\binom{3+2+3}{2} = 28$.

Conditional distribution

$$P_{X,Y}(x,y) = \frac{\text{ways to choose } x \text{ blue pens} \quad \text{ways to choose } y \text{ red pens} \quad \text{ways to choose } 2-x-y \text{ green pens}}{\binom{3}{x} \binom{2}{y} \binom{3}{2-x-y}} \quad \text{for } x, y \in \{0, 1, 2\}, \\ x+y \leq 2.$$

$x \backslash y$	0	1	2	P_Y
0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
1	$\frac{6}{28}$	$\frac{6}{28}$	0	$\frac{12}{28}$
2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
P_X	$\frac{10}{28}$	$\frac{15}{28}$	$\frac{3}{28}$	

$$(b) P((X,Y) \in A) = P(X+Y \leq 1) = P_{X,Y}(0,0) + P_{X,Y}(0,1) + P_{X,Y}(1,0) \\ = \frac{3}{28} + \frac{6}{28} + \frac{9}{28} = \frac{18}{28}.$$

Conditional distribution

(c) Note that $P_Y(1) = 12/28 = P(Y=1)$

Hence, $P(X=0|Y=1) = \frac{P(X=0, Y=1)}{P(Y=1)} = \frac{6/28}{12/28} = \frac{1}{2}$.

Similarly, verify that (Homework): $P(X=1|Y=1) = \frac{1}{2}$,
 $P(X=2|Y=1) = 0$.

$$\begin{aligned}(d) F_{X|Y=1}(1) &= P(X \leq 1 | Y=1) = \frac{P(X \leq 1, Y=1)}{P(Y=1)} \\&= \frac{P(X=0, Y=1) + P(X=1, Y=1)}{P(Y=1)} \\&= \frac{\frac{6/28 + 6/28}{12/28}}{1} = 1.\end{aligned}$$

Conditional distribution

- ▶ Wireless communication: In practice, we use channels (e.g., mobile communication via a wireless channel) to communicate information. But in the physical world, the channels are usually not “perfect”. That is, due to the noise in the channels, a transmitted message may be received as some other message.
- ▶ Example: The input messages to a channel are chosen from the set $\{0, 1\}$ with probability $P(X = 0) = .5$ and $P(X = 1) = .5$. Output of the channel is a stream of messages from the set $\{0, 1\}$ with probability $P(Y = 0)$ and $P(Y = 1)$.
- ▶ In the channel, the input message 0 is altered to 1 with probability p and the input message 1 is altered to 0 with probability q .

Conditional distribution

► Hence,

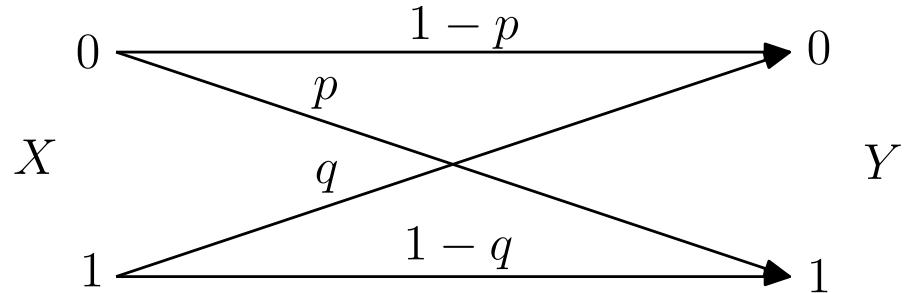
$$P(Y = 0|X = 0) = 1 - p, \quad (1)$$

$$P(Y = 1|X = 0) = p, \quad (2)$$

$$P(Y = 1|X = 1) = 1 - q, \quad (3)$$

$$P(Y = 0|X = 1) = q. \quad (4)$$

► The channel is depicted in the following figure.



► If $p = .25, q = .35$, what is the PMF of Y ?

Conditional distribution

Solⁿ:

$$\begin{aligned} p_Y(0) &= P_{X,Y}(0,0) + P_{X,Y}(1,0) \\ &= P(X=0, Y=0) + P(X=1, Y=0) \\ &= P(Y=0|X=0) \cdot P(X=0) + P(Y=0|X=1) \cdot P(X=1) \\ &= (1-p) \times 0.5 + q \times 0.5 \\ &= 0.75 \times 0.5 + 0.35 \times 0.5 \\ &= 0.55 \end{aligned}$$

Similarly,

$$\begin{aligned} p_Y(1) &= P(Y=1|X=0) \cdot P(X=0) + P(Y=1|X=1) \cdot P(X=1) \\ &= [p + (1-q)] \times 0.5 \\ &= 0.45. \end{aligned}$$

Alternatively,

$$\begin{aligned} p_Y(1) &= P(Y=1) \\ &= 1 - P(Y=0) \quad \left(\because Y \text{ can take either } 0 \text{ or } 1 \text{ value} \right) \\ &= 1 - 0.55 \\ &= 0.45. \end{aligned}$$
$$\Rightarrow P(Y=1) = 1 - P(Y \neq 1)$$
$$= 1 - P(Y=0)$$

Lecture 14: Expectation and Variance - Part I

Satyajit Thakor
IIT Mandi

31 March, 2020

Expectation

- ▶ Let X be a random variable describing the amount won in a game.
- ▶ A question is: How much does a person win “on average” in a game?
- ▶ For example, Let X be the amount won and is equal to the outcome of a fair dice. Then how much does a person win “on average”? or What is the “expected value” of winning amount?

$$\frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5.$$

- ▶ This is the same as the arithmetic mean of n real numbers x_1, \dots, x_n defined as

$$\frac{x_1 + \dots + x_n}{n}.$$

- ▶ If the dice is biased (not fair) then how much a person wins “on average” in a game?

Expectation

- If X is a discrete r.v. taking on the possible values x_1, x_2, \dots , then the expectation or expected value or mean of X , denoted by $E(X)$, is defined as

$$E(X) = \sum_i x_i P(X = x_i).$$

- Example: If the dice is biased (not fair) such that

$$P(X = 1) = 0, P(X = 2) = P(X = 3) = P(X = 4) = 1/6,$$

$$P(X = 5) = P(X = 6) = 1/4,$$

then how much a person wins “on average” in a game?

$$\begin{aligned} E(X) &= 0 \cdot 1 + (2+3+4) \cdot \frac{1}{6} + (5+6) \cdot \frac{1}{4} \\ &= \frac{3}{2} + \frac{11}{4} = \frac{17}{4} = 4.25. \end{aligned}$$

Expectation

- ▶ Example: Find $E(X)$ if $X \sim \text{Bern}(p)$.

$$E(X) = 1 \cdot p + 0 \cdot (1-p) = p.$$

- ▶ Example: What is the expected value of a binomially distributed r.v.?

- Recall: $X \sim \text{Bin}(n, p)$ if

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k=0, 1, \dots, n$$

otherwise.
 $= 0$

- We will use the following identity:

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

Expectation

$$\begin{aligned} - \text{Now, } E(X) &= \sum_{k=0}^n k P(X=k) \\ &= \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} \\ &= n \sum_{k=0}^n \binom{n-1}{k-1} p^k (1-p)^{n-k} \quad (\because \text{using the identity}) \\ &= np \sum_{k=1}^n \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\ &= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} \\ &\quad \underbrace{\text{PMF of } \text{Bin}(n-1, p)}_{\text{for } j=0, 1, \dots, n-1.} \\ &= np. \quad \text{PMF always sums up to 1} \end{aligned}$$

Properties of expectation

- Expectation of a function of r.v.: If X is a discrete r.v. and $g : \mathbb{R} \rightarrow \mathbb{R}$ is a function, then

$$E(g(X)) = \sum_x g(x)P(X=x) = \sum_x g(x)p_X(x)$$

where the sum is taken over all possible values of X .

- How?: Recall that an r.v. X is a function $X : S \rightarrow \mathbb{R}$.

Hence,

$$\begin{aligned} E(g(X)) &= \sum_{s \in S} g(X(s)) P(\{s\}) \\ &= \sum_{x \in \mathbb{R}} \sum_{s: X(s)=x} g(X(s)) P(\{s\}) \\ &= \sum_{x \in \mathbb{R}} g(x) \sum_{s: X(s)=x} P(\{s\}) \\ &= \sum_x g(x) P(X=x). \end{aligned}$$

Properties of expectation

- ▶ Similarly, for a function of two r.v.s, we have

$$\begin{aligned} E(g(X, Y)) &= \sum_x \sum_y g(x, y) P(X = x, Y = y) \\ &= \sum_x \sum_y g(x, y) p_{X,Y}(x, y). \end{aligned}$$

- ▶ Suppose X has the following PMF

$$p_X(0) = .2, p_X(1) = .5, p_X(2) = .3.$$

Find $E[X^2]$.

Let $Y = g(X) = X^2$. Then,

$$\begin{aligned} E(X^2) = E(Y) &= 0^2 \cdot (0.2) + 1^2 \cdot (0.5) + 2^2 \cdot (0.3) \\ &= 1.7. \end{aligned}$$

Properties of expectation

- Linearity: Expectation is linear, i.e., for r.v.s X and Y and a constant c ,

1. $E(X + Y) = E(X) + E(Y)$, Let $g(X, Y) = X + Y$.
2. $E(cX) = cE(X)$.

$$\begin{aligned} E(X+Y) &= E(g(X+Y)) = \sum_x \sum_y g(x, y) p_{X,Y}(x, y) \\ &= \sum_x \sum_y (x+y) p_{X,Y}(x, y) \\ &= \sum_x \sum_y x p_{X,Y}(x, y) + \sum_x \sum_y y p_{X,Y}(x, y) \\ &= \sum_x \sum_y x p_{X,Y}(x, y) + \sum_y \sum_x y p_{X,Y}(x, y) \\ &= \sum_x x p_X(x) + \sum_y y p_Y(y) \\ &= E(X) + E(Y) \end{aligned}$$

$$\begin{aligned} E(cX) &= \sum_x c x p_X(x) \\ &= c \sum_x x p_X(x) \\ &= c E(X). \end{aligned}$$

Properties of expectation

- ▶ Example: A construction firm has sent in bids for 3 jobs worth (in profits) 10, 20, and 40 (lakh) rupees. If its probabilities of winning the jobs are respectively .2, .8, and .3, what is the firm's expected total profit?

- Let X_i be the profit from job i , $i \in \{1, 2, 3\}$.
- Total profit = $X_1 + X_2 + X_3$.
- Expected total profit = $E(X_1 + X_2 + X_3)$
= $E(X_1) + E(X_2) + E(X_3)$
= $10 \cdot 0.2 + 20 \cdot 0.8 + 40 \cdot 0.3$
= $2 + 16 + 12$
= 30 Lakh rupees.

Lecture 15: Expectation and Variance - Part II

Satyajit Thakor
IIT Mandi

1 April, 2020

Properties of expectation

- If X and Y are independent r.v.s, then

$$E(XY) = E(X)E(Y).$$

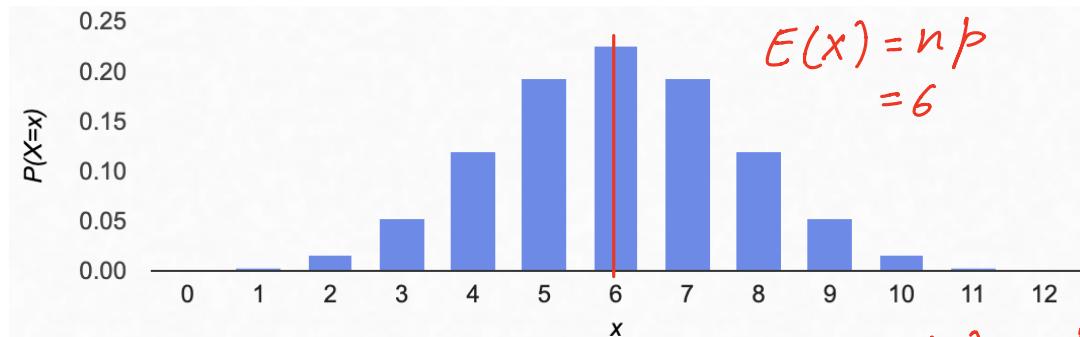
Proof:

$$\begin{aligned} E(XY) &= \sum_x \sum_y xy p_{XY}(x,y) \\ &= \sum_x \sum_y xy p_X(x) p_Y(y) \\ &= \sum_x x p_X(x) \sum_y y p_Y(y) \\ &= E(X) E(Y) \end{aligned}$$

Expectation

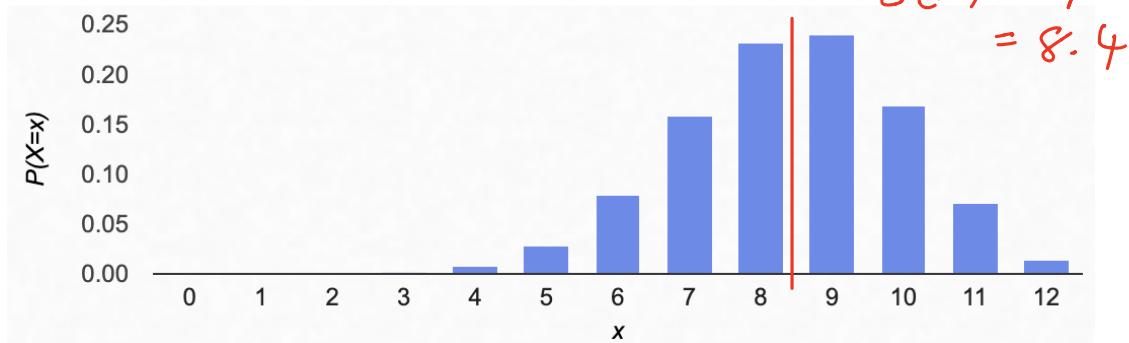
- Expectation of r.v.s with binomial distribution.

$\text{Bin}(12, 0.5)$



$$\begin{aligned}E(X) &= np \\&= 6\end{aligned}$$

$\text{Bin}(12, 0.7)$



$$\begin{aligned}E(X) &= np \\&= 8.4\end{aligned}$$

- The distribution is “centred around” its expected value.

Expectation Recall: $X \sim \text{Geom}(p)$ if $P(X=k) = (1-p)^k \cdot p$
 for $k=0, 1, 2, \dots$

- What is the expected value of a random variable with geometric distribution?

$$\begin{aligned}
 E(X) &= \sum_{k=0}^{\infty} k \cdot p \cdot (1-p)^k \\
 &= p \sum_{k=1}^{\infty} k (1-p)^k \\
 &= p(1-p) \sum_{k=1}^{\infty} k (1-p)^{k-1} \\
 &= p(1-p) \frac{1}{p^2} \\
 &= \frac{1-p}{p}
 \end{aligned}$$

∴ Recall: for $0 < x < 1$,

$$\sum_{k=1}^{\infty} x^k = \frac{x}{1-x}$$

Take derivative:

$$\frac{d}{dx} \sum_{k=1}^{\infty} x^k = \frac{d}{dx} \underbrace{\frac{x}{1-x}}_{\text{Homework}}$$

$$\Rightarrow \sum_{k=1}^{\infty} k x^{k-1} = \frac{1}{(1-x)^2}$$

Variance

- ▶ Recall: Expected value of a random variable is a single value.
- ▶ It tells us the center of mass of the distribution of an r.v.
- ▶ That is, the distribution is “centred around” its mean value.
- ▶ Variance of a random variable is also a single value.
- ▶ It tells us how “spread out” the distribution is.
- ▶ The variance of an r.v. X is

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

- ▶ The square root of the variance is called the standard deviation:

$$\text{SD}(X) = \sqrt{\text{Var}(X)}.$$

Variance

- ▶ $E(X)$: Mean of an r.v. X is often denoted by μ_X .
- ▶ $\text{Var}(X)$: Variance of an r.v. X is often denoted by σ_X .
- ▶ Theorem: For any r.v. X ,

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

Proof: $\text{Var}(X) = E(X - E(X))^2$

$$\begin{aligned} &= E(X - \mu_X)^2 \\ &= E(X^2 - 2\mu_X X + \mu_X^2) \\ &= E(X^2) - 2\mu_X E(X) + \mu_X^2 \quad (\because \text{Linearity of expectation}) \\ &= E(X^2) - 2\mu_X^2 + \mu_X^2 \\ &= E(X^2) - \mu_X^2 \\ &= E(X^2) - E(X)^2 \end{aligned}$$

Expectation of a constant is the constant
Let $g(x) = c$. Then
 $E(c) = \sum_x c p_X(x)$
 $= c.$

Variance

- ▶ Example: Find $\text{Var}(X)$ if $X \sim \text{Bern}(p)$.

- In Lecture 14, we showed that $E(X) = p$.

$$\begin{aligned}\text{Var}(X) &= E(X^2) - E(X)^2 \\ &= \sum_x x^2 p_x(x) - p^2 \\ &= 1^2 \cdot p + 0^2 \cdot (1-p) - p^2 \\ &= p - p^2 \\ &= p(1-p).\end{aligned}$$

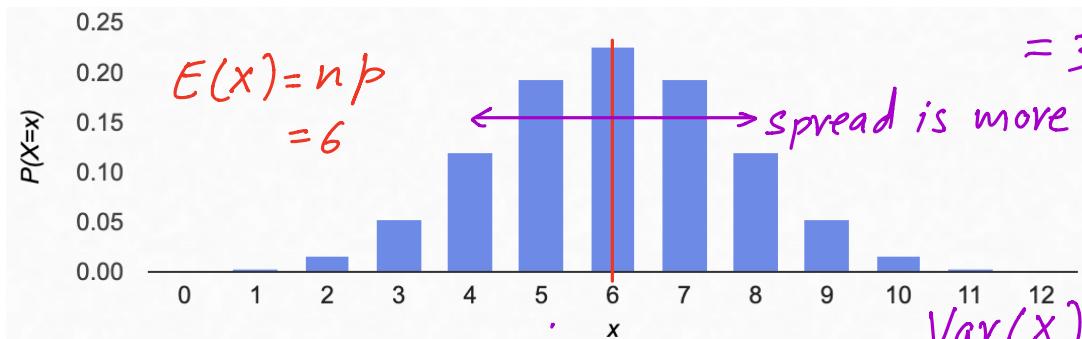
- ▶ Assignment problem: show that the variance of a binomially distributed r.v. $X \sim \text{Bin}(n, p)$ is

$$\text{Var}(X) = np(1-p)$$

Variance

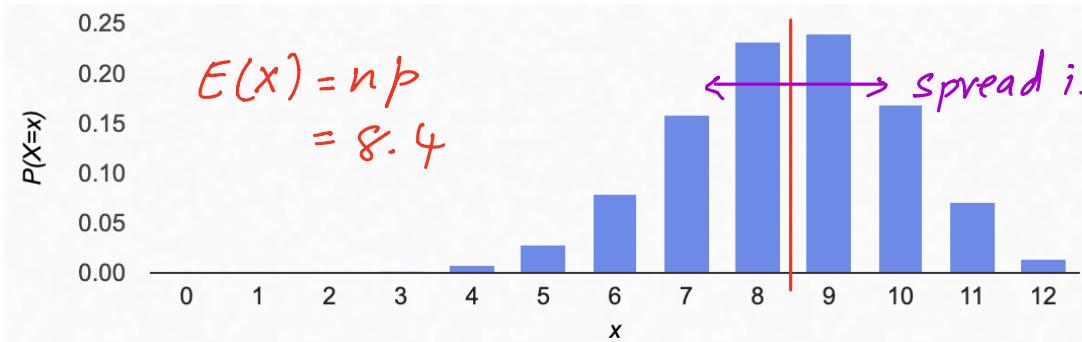
- Variance of r.v.s with binomial distribution. $\text{Var}(X) = np(1-p)$

$\text{Bin}(12, 0.5)$



$$\text{Var}(X) = np(1-p) = 3$$

$\text{Bin}(12, 0.7)$



$$\text{Var}(X) = 2.52$$

- The distribution is “centred around” its expected value.
- Variance tells us how “spread out” the distribution is.

Lecture 16: Expectation and Variance - Part III

Satyajit Thakor
IIT Mandi

3 April, 2020

Properties of variance

- $\text{Var}(X) \geq 0$.

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

$$= \sum_x (x - \mu_X)^2 p_X(x)$$

$$\geq 0 \quad (\because (x - \mu_X)^2 \geq 0, p_X(x) \geq 0 \text{ for all } x)$$

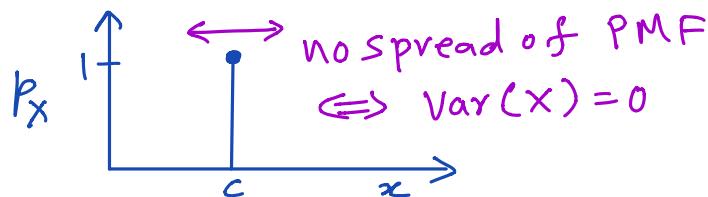
- $\text{Var}(c) = 0$ for any constant c .

Let X be an r.v. s.t. $\underbrace{P(X=c)=1}_{\text{square of a real no.}} \Rightarrow E(X) = \sum_x x P(X=x) = c \cdot P(X=c)$

Then, for r.v. $(X-c)^2$, $\underbrace{P((X-c)^2=0)}_{0 \cdot 1} = 1$.

$$\Rightarrow E[(X-c)^2] = 0 \cdot 1 = 0$$

Note: $\text{Var}(c) = 0$ means that when the r.v. X is a constant c (i.e., $P(X=c)=1$) then the spread of PMF is zero:



Properties of variance

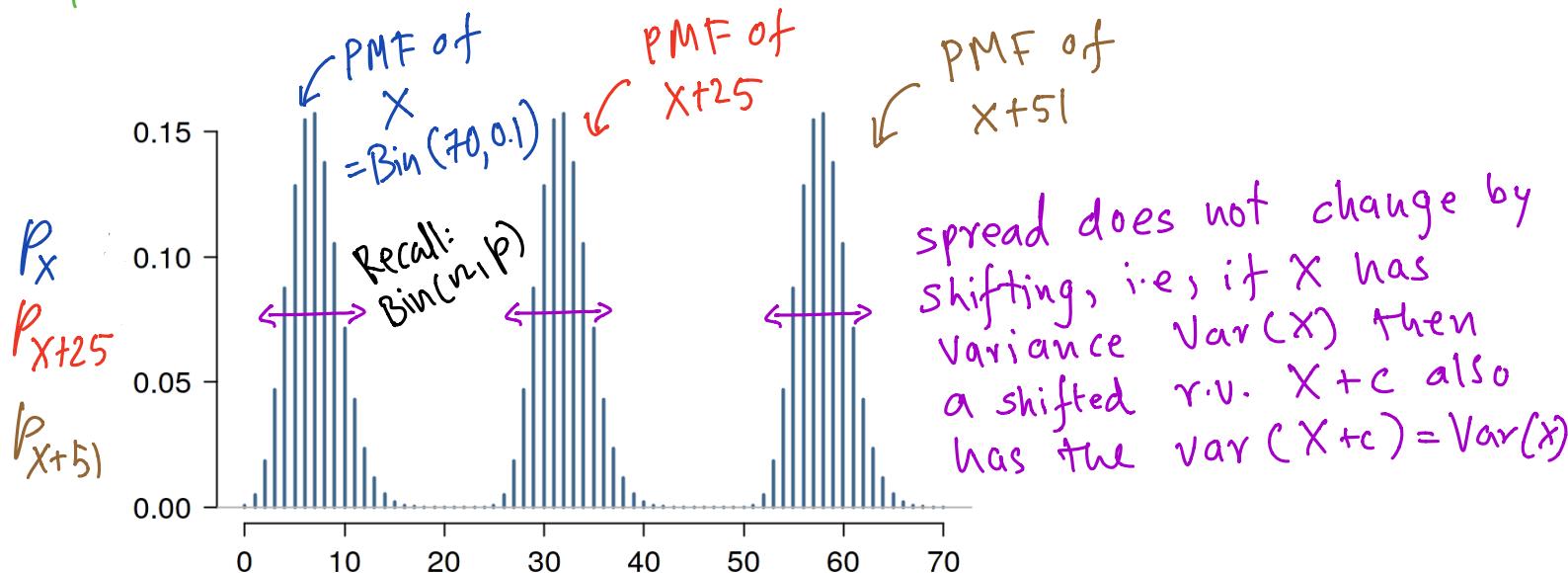
shifting X by c

- $\text{Var}(X + c) = \text{Var}(X)$ for any constant c .

- Recall that $E(X+c) = \mu_x + c$.

- But $\text{Var}(X+c) = E[(X+c - E(X+c))^2] = E[(X+c - \mu_x - c)^2] = E[(X - \mu_x)^2] = \text{Var}(X)$.

- Note: spread does not change by "shifting".



Properties of variance

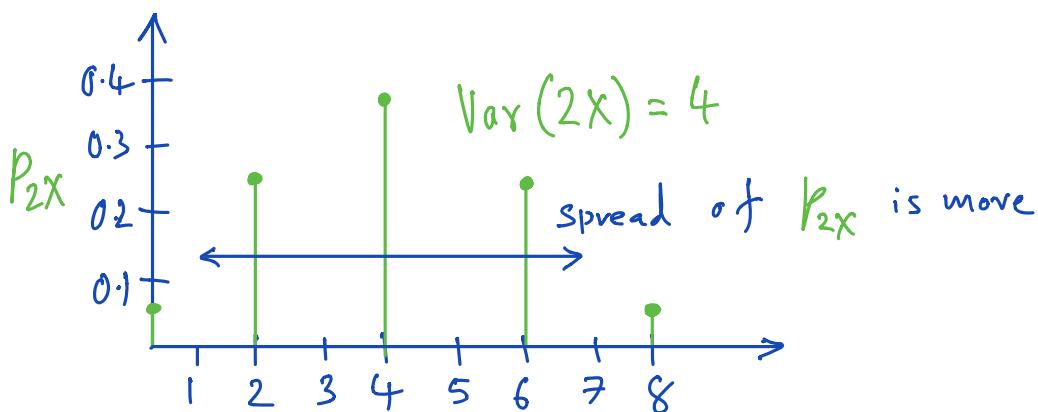
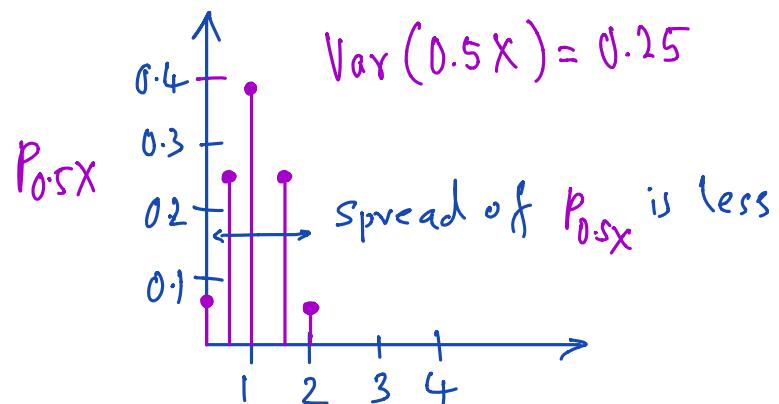
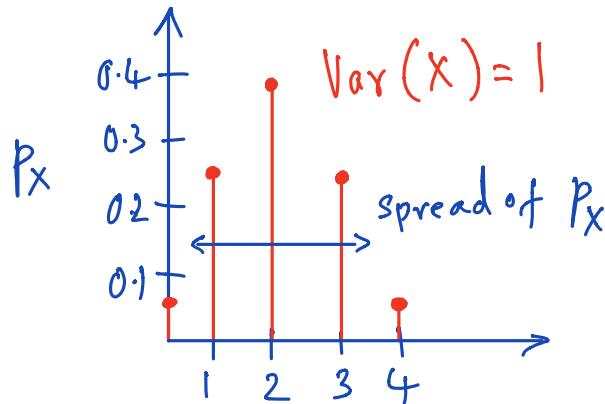
- ▶ $\text{Var}(cX) = c^2 \text{Var}(X)$ for any constant c .

$$\begin{aligned}\text{Var}(cX) &= E[(cX - E(cX))^2] \\ &= E[c^2(X - \mu_X)^2] \\ &= c^2 E[(X - \mu_X)^2] \quad \because \text{linearity of } E(\cdot). \\ &= c^2 \text{Var}(X).\end{aligned}$$

- That is, if we scale X by c then the variance $\text{Var}(cX)$ of the resulting r.v. cX is $\text{Var}(X)$ scaled by c^2 .
- So, depending on the value of c , the spread for cX either reduces or increases (if $c=1$ then the spread remains the same.)
- The next example demonstrates this.

Properties of variance

- Example: Let $X \sim \text{Bin}(4, .5)$. Plot PMFs of X , $2X$, $.5X$ and compare their variance.



Functions of independent random variables

Theorem:

- Suppose that X and Y are independent r.v.s. Then $W = g(X)$ and $Z = h(Y)$ are also independent.

Proof:

$$\begin{aligned} P_{W,Z}(w,z) &= P(W=w, Z=z) = P\left(\left[\bigcup_{x: g(x)=w} \{X=x\}\right] \cap \left[\bigcup_{y: h(y)=z} \{Y=y\}\right]\right) \\ &= P\left(\bigcup_{\substack{x: g(x)=w \\ y: h(y)=z}} \{X=x, Y=y\}\right) \\ &= \sum_{x: g(x)=w} \sum_{y: h(y)=z} P_{XY}(x,y) \\ &\quad \Downarrow \text{since } P_{XY}(x,y) = P_X(x) \cdot P_Y(y) \\ &= \sum_{x: g(x)=w} P_X(x) \sum_{y: h(y)=z} P_Y(y) = P_W(w) \cdot P_Z(z). \end{aligned}$$

Note: You may skip the proof (will not be asked in exam). But remember the theorem statement.

The proof is just to satisfy your curiosity.

Properties of variance

- If X and Y are independent, then

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

$$\begin{aligned}\text{Var}(X+Y) &= E[(X+Y - E(X+Y))^2] \\ &= E[(X+Y - \mu_X - \mu_Y)^2] \quad \because \text{linearity of } E(\cdot) \\ &= E[(X-\mu_X)^2 + (Y-\mu_Y)^2 - 2(X-\mu_X)(Y-\mu_Y)] \\ &= \text{Var}(X) + \text{Var}(Y) - 2E[\underbrace{(X-\mu_X)(Y-\mu_Y)}_{\substack{\downarrow \\ \text{functions of independent} \\ \text{r.v.s are independent}}}] \\ &= \text{Var}(X) + \text{Var}(Y) - 2E(X-\mu_X) \cdot E(Y-\mu_Y) \\ &= \text{Var}(X) + \text{Var}(Y) - 2\underbrace{(\mu_X - \mu_X)}_0 \cdot \underbrace{(\mu_Y - \mu_Y)}_0 \\ &= \text{Var}(X) + \text{Var}(Y)\end{aligned}$$

Example

► Let X be a random variable with the PMF

$$p_X(-3) = 1/6, p_X(6) = 1/2, p_X(9) = 1/3.$$

- (a) Find $\mu_{g(X)} = E(g(X))$, where $g(X) = (2X + 1)^2$.
(b) Find $\text{SD}(g(X))$.

$$\begin{aligned} \text{(a)} \quad \mu_{g(X)} &= E[(2X+1)^2] \\ &= \frac{25}{16} + \frac{169}{2} + \frac{361}{3} = 209. \end{aligned}$$

$$\begin{aligned} \text{(b)} \quad \sigma_{g(X)}^2 &= E[(2X+1)^2 - 209]^2 \\ &= \sum_x [(2x+1)^2 - 209] \cdot p_X(x) \\ &= (25 - 209)^2 \cdot \frac{1}{6} + (169 - 209)^2 \cdot \frac{1}{2} + (361 - 209)^2 \cdot \frac{1}{3} \\ &= 14144 \quad \Rightarrow \quad \text{SD}(g(X)) = \sqrt{\sigma_{g(X)}^2} \\ &= 118.9. \end{aligned}$$

Lecture 17: Expectation and Variance - Part IV & Continuous Random Variables - Part I

Satyajit Thakor
IIT Mandi

6 April, 2020

Example

- ▶ There are k distinguishable balls and n distinguishable boxes. The balls are randomly placed in the boxes, with all n^k possibilities equally likely. Problems in this setting are called occupancy problems, and are used in computer science (for example, for randomized algorithms).
- ▶ Find the expected number of empty boxes in terms of the parameters n, k .
 - Let X_i be the r.v. such that $\{X_i=1\}$ corresponds to i th box being empty and $\{X_i=0\}$ corresponds to i th box not empty.
 - Then, the no. of empty boxes is the r.v. $X = \sum_{i=1}^n X_i$.
 - Expected no. of empty boxes is

Example

$$E(X) = E \sum_{i=1}^n X_i = \sum_{i=1}^n E(X_i) \quad (\because \text{linearity of } E(\cdot))$$

-Now, $E(X_i) = P(X_i=1) \cdot 1$

$$= P(\{\text{there are no balls in } i\text{th box}\})$$

$$= P\left(\bigcap_{j=1}^K \underbrace{\{\text{j}^{th} \text{ ball is not placed in } i\text{th box}\}}_{\Downarrow \because \text{independent events}}\right)$$

$$= \prod_{j=1}^K P(\{\text{j}^{th} \text{ ball is not placed in } i\text{th box}\})$$

$$= \prod_{j=1}^K \left[1 - P(\{\text{j}^{th} \text{ ball is placed in } i\text{th box}\}) \right] \quad \begin{matrix} \text{underbrace} \\ \because \text{ball is placed in any of } n \text{ boxes randomly.} \end{matrix}$$

$$= \prod_{j=1}^K \left[1 - \frac{1}{n} \right] = \left[1 - \frac{1}{n} \right]^K.$$

Example

- Suppose that X and Y are independent r.v.s for which $\text{Var}(X) = \text{Var}(Y) = 3$. Find $\text{Var}(X - Y)$.
- X and Y are independent $\Rightarrow X$ and $-Y$ are independent.

(\because functions of independent r.v.s are independent)

- Now,

$$\begin{aligned}\text{Var}(X - Y) &= \text{Var}(X + (-Y)) \\ &= \text{Var}(X) + \text{Var}(-Y) \quad \left(\begin{array}{l} \because z \text{ and } w \text{ are independent} \\ \Rightarrow \text{Var}(z+w) = \text{Var}(z) + \text{Var}(w) \end{array} \right)\end{aligned}$$

$$= \text{Var}(X) + (-1)^2 \text{Var}(Y) \quad \left(\because \text{Var}(cX) = c^2 \text{Var}(X) \right)$$

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

$$= 3 + 3 = 6 .$$

Example

- A person wishes to insure his car for 200,000 rupees. The insurance company estimates that a total loss¹ will occur with probability 0.002, a 50% loss with probability 0.01, and a 25% loss with probability 0.1. Ignoring all other partial losses, what premium should the insurance company charge each year to realize an average profit of 500 rupees?

- Let X = claim amount. Then,

$x:$	200,000	100,000	50,000	0
$p_x(x):$	0.002	0.01	0.1	0.888

$$\Rightarrow \text{Expected claim} = E(X) = 200000 \cdot 0.002 + 100000 \cdot 0.01 + 50000 \cdot 0.1$$

required
average profit

$$= 6400/-$$

Hence, the company should charge the premium : $6400 + 500$
 $= 6900/-$

¹Here, loss of the insurance company means the claim of the person after accident of the car. This amount the company will pay to the person.

Example

- A coin is biased such that a head is three times as likely to occur as a tail. Find the expected number of tails when this coin is tossed twice.

- For the given biased coin : $P(\{H\}) = \frac{3}{4}$, $P(\{T\}) = \frac{1}{4}$.

- Let X be the no. of tails in two tosses. Then

$$P(X=0) = P(\{H, H\}) = \frac{9}{16}$$

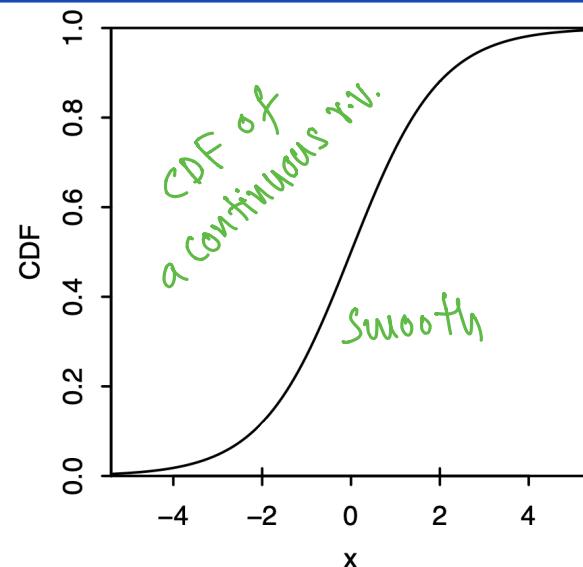
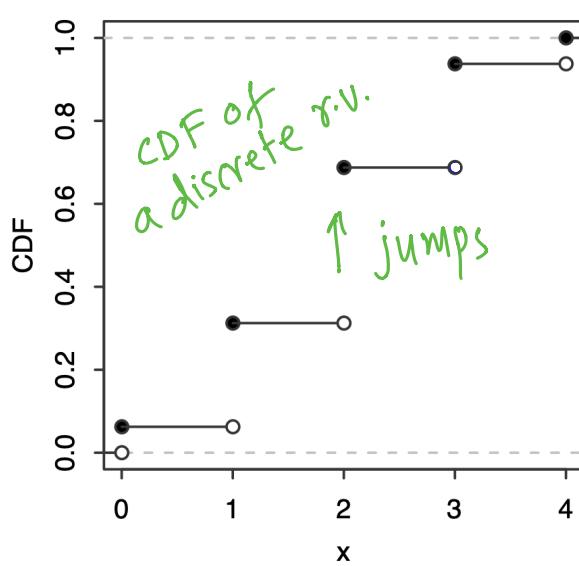
$$P(X=1) = P(\{H, T\}) + P(\{T, H\}) = 2 \cdot \frac{3}{4} \cdot \frac{1}{4} = \frac{3}{8}$$

$$P(X=2) = P(\{T, T\}) = \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16}$$

$$\text{Then, } E(X) = 0 \cdot \frac{9}{16} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{1}{16} = \frac{4}{8} = \frac{1}{2}.$$

Continuous random variables

- ▶ An r.v. has a **continuous distribution** if its CDF is differentiable.
(or if it is at least (1) continuous and (2) not differentiable at finite number of points)
- ▶ A **continuous r.v.** is an r.v. with a continuous distribution.



- ▶ CDF of a discrete r.v. has jumps, CDF of a continuous r.v. is smooth.

Continuous random variables

- ▶ For a continuous r.v. X with CDF F_X , the probability density function (PDF) of X is the derivative f_X of the CDF, given by

$$f_X(x) = F'_X(x) = \frac{d}{dx}F_X(x).$$

- ▶ PDF to CDF: Since F_X is an antiderivative of f_X , CDF can be obtained by integration of PDF:

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

- ▶ Caution:
 1. For a continuous r.v. X , $P(X = x) = 0$ for all x because $P(X = x)$ is the height of a jump in the CDF at x , but the CDF of X has no jumps.
 2. $f_X(x)$ is not a probability, and in fact it is possible to have $f_X(x) > 1$ for some values of x (example will be discussed in the next lecture).

Lecture 18: Continuous Random Variables - Part II

Satyajit Thakor
IIT Mandi

7 April, 2020

Continuous random variables

- ▶ Continuous random variables can take any value within an interval (or continuous region).
- ▶ Example: Let X be the time to failure of a newly charged battery.
- ▶ Failure can be defined to be the moment at which the battery can no longer supply enough energy to operate a certain appliance.
- ▶ The r.v. X is continuous since it can take any positive value, i.e., any value in the interval $[0, \infty)$.

- Recall: $f_X(x) = \frac{d}{dx} F_X(x)$ (compute PDF from CDF)

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \quad (\text{compute CDF from PDF})$$

Continuous random variables

- ▶ By definition of CDF,

$$P(X \leq x) = F_X(x) = \int_{-\infty}^x f_X(t)dt.$$

- ▶ Similarly, the probability that X takes values in the interval $[a, b]$ or $(a, b]$ or $[a, b)$ or (a, b) is

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

$$= F_X(b) - F_X(a) = \int_a^b f_X(x)dx.$$

- ▶ Also,

$$P(X = x) = \int_x^x f_X(t)dt = 0$$

area under the PDF curve from a to b.

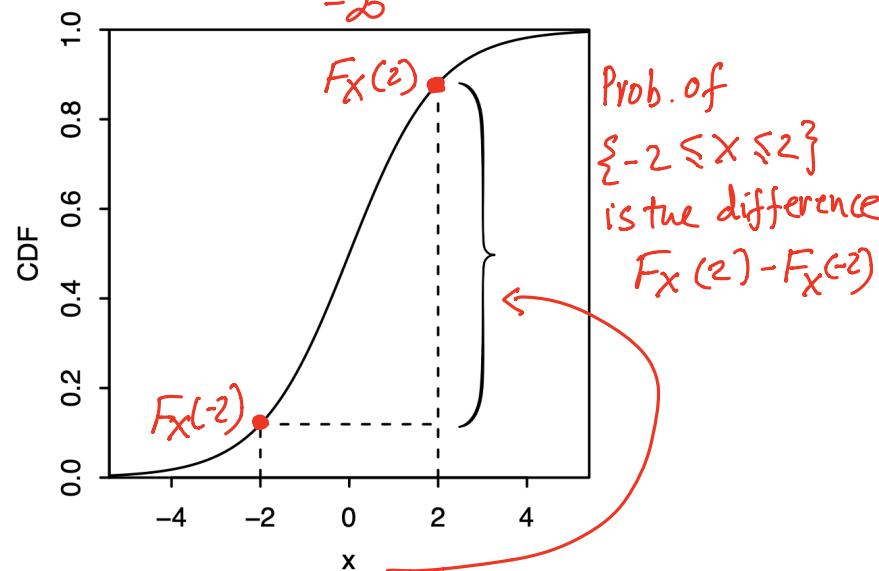
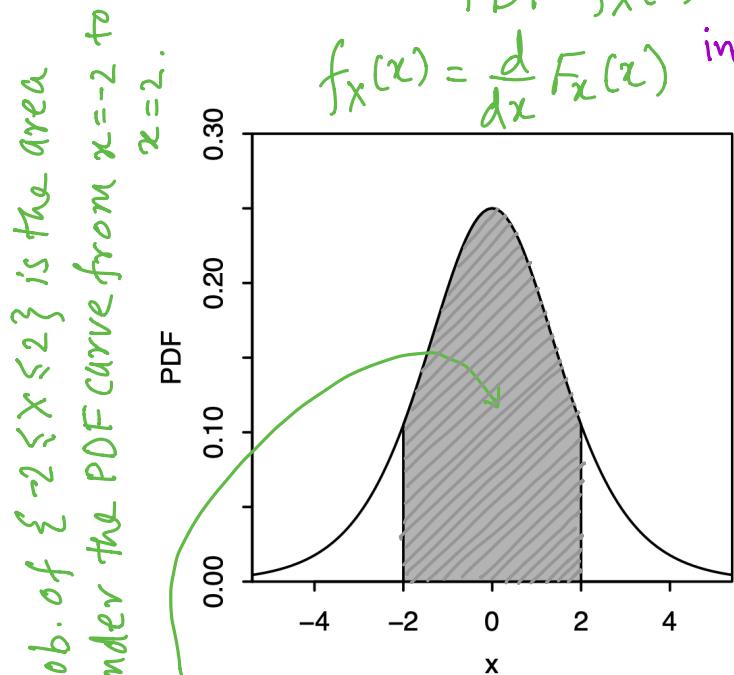
for a continuous r.v. X , $P(X=x)=0$ for all $x \in \mathbb{R}$.

Continuous random variables

- ▶ Example: PDF and CDF differentiate

$$\text{PDF } f_X(x) \longleftrightarrow \text{CDF } F_X(x)$$

$$f_X(x) = \frac{d}{dx} F_X(x) \text{ integrate } F_X(x) = \int_{-\infty}^x f_X(t) dt$$



$$P(-2 \leq X \leq 2) = \int_{-2}^2 f_X(x) dx$$

$$P(-2 \leq X \leq 2) = F_X(2) - F_X(-2)$$

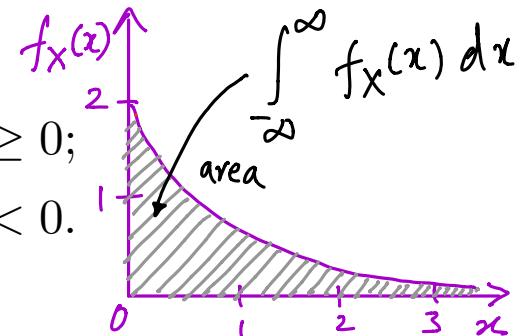
Prob. of $\{-2 \leq X \leq 2\}$ is the difference $F_X(2) - F_X(-2)$

Continuous random variables

That is, the area under the PDF curve from $-\infty$ to ∞ is 1.

- ▶ Recall: a valid PMF must be nonnegative and sum to 1.
- ▶ Similarly, a valid PDF must be nonnegative and integrate to 1.
- ▶ Example: Suppose that the battery failure time, measured in hours, has the PDF

$$f_X(x) = \begin{cases} f(x) = \frac{2}{(x+1)^3}, & x \geq 0; \\ 0, & x < 0. \end{cases}$$



- ▶ (a) Is this a valid PDF?

(1) Check "non-negativity": Yes, $f_X(x)$ is non-negative for all $x \in \mathbb{R}$

(2) check "integrate to 1":

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_0^{\infty} \frac{2}{(x+1)^3} dx = -\frac{1}{(x+1)^2} \Big|_{x=0}^{x=\infty} = 0 - (-1) = 1$$

From (1) & (2) the PDF is valid.

$$\begin{aligned} \text{Let } u &= x+1 \\ \Rightarrow du &= dx \\ \Rightarrow \int_0^{\infty} \frac{2}{(x+1)^3} dx &= \int_1^{\infty} \frac{2}{u^3} du \\ &= 2 \cdot \frac{1}{2u^2} \Big|_1^{\infty} = \frac{1}{(x+1)^2} \Big|_0^{\infty} \end{aligned}$$

Continuous random variables

Prob. that battery lasts no more than 5 hours.

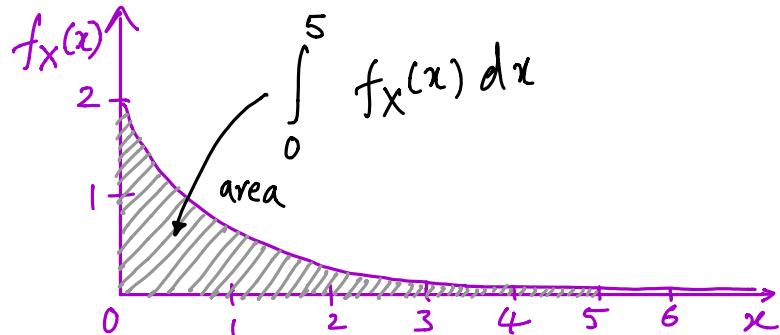
- (b) Find $P(X \in [0, 5])$ using the PDF.

$$P(X \in [0, 5]) = P(0 \leq X \leq 5)$$

$$= \int_0^5 \frac{2}{(x+1)^3} dx$$

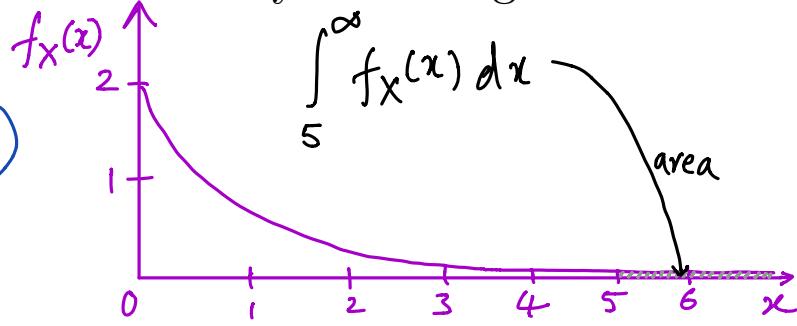
$$= -\frac{1}{(x+1)^2} \Big|_0^5$$

$$= -\frac{1}{36} - \frac{-1}{1} = \frac{35}{36}.$$



- (c) What is the probability that the battery lasts longer than 5 hours?

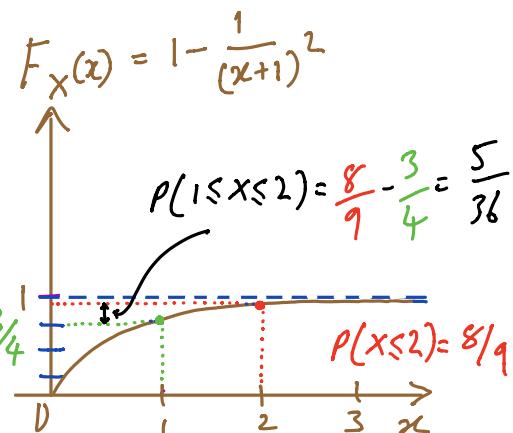
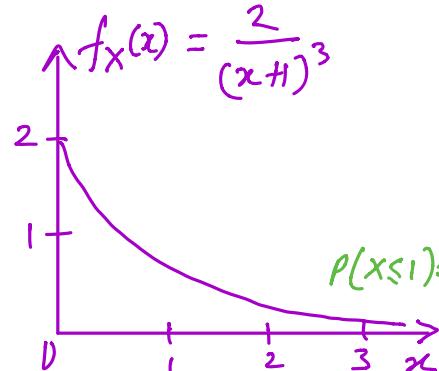
$$\begin{aligned} P(X > 5) &= 1 - P(X \leq 5) \\ &= 1 - P(0 \leq X \leq 5) \\ &= 1 - \frac{35}{36} = \frac{1}{36} \end{aligned}$$



Continuous random variables

- (d) Find the CDF of the battery failure time.

$$\begin{aligned}
 F_X(x) &= P(X \leq x) \\
 &= \int_0^x \frac{2}{(t+1)^3} dt \\
 &= \left[-\frac{1}{(t+1)^2} \right]_{t=0}^{t=x} \\
 &= \frac{-1}{(x+1)^2} - (-1) = 1 - \frac{1}{(x+1)^2}
 \end{aligned}$$



- (e) Find $P(X \in [1, 2])$ using the CDF.

$$\begin{aligned}
 P(X \in [1, 2]) &= P(1 \leq X \leq 2) \\
 &= F_X(2) - F_X(1) \\
 &= \frac{8}{9} - \frac{3}{4} = \frac{5}{36}
 \end{aligned}$$

Lecture 19: Continuous Random Variables - Part III

Satyajit Thakor
IIT Mandi

10 April, 2020

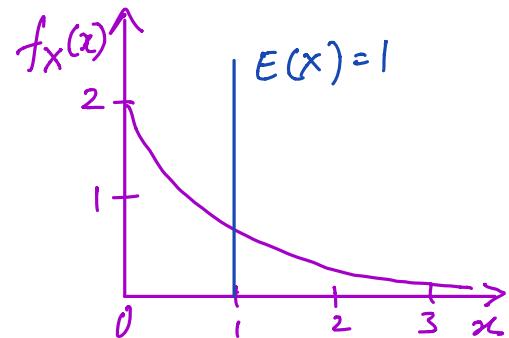
Continuous random variables

- The expected value or expectation or mean of a continuous r.v. with a probability density function $f_X(x)$ is

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

- Example continued:(f) Find the expected battery failure time.

$$\begin{aligned} E(X) &= \int_0^{\infty} \left[x \frac{2}{(x+1)^3} \right] dx \\ &= \int_0^{\infty} \left[\frac{2}{(x+1)^2} - \frac{2}{(x+1)^3} \right] dx \\ &= \left. \frac{-2}{(x+1)} + \frac{1}{(x+1)^2} \right|_{x=0}^{x=\infty} \\ &= 0 - (-1) = 1. \end{aligned}$$



Continuous random variables

- The median of a continuous r.v. X with a CDF $F_X(x)$ is the value x such that $F_X(x) = 0.5$.
- The r.v. is equally likely to fall above or below the median value.
- Example continued:(g) Find the median of the battery failure time.

We need to solve $F_X(x) = 0.5$.

$$\Rightarrow 1 - \frac{1}{(x+1)^2} = \frac{1}{2}$$

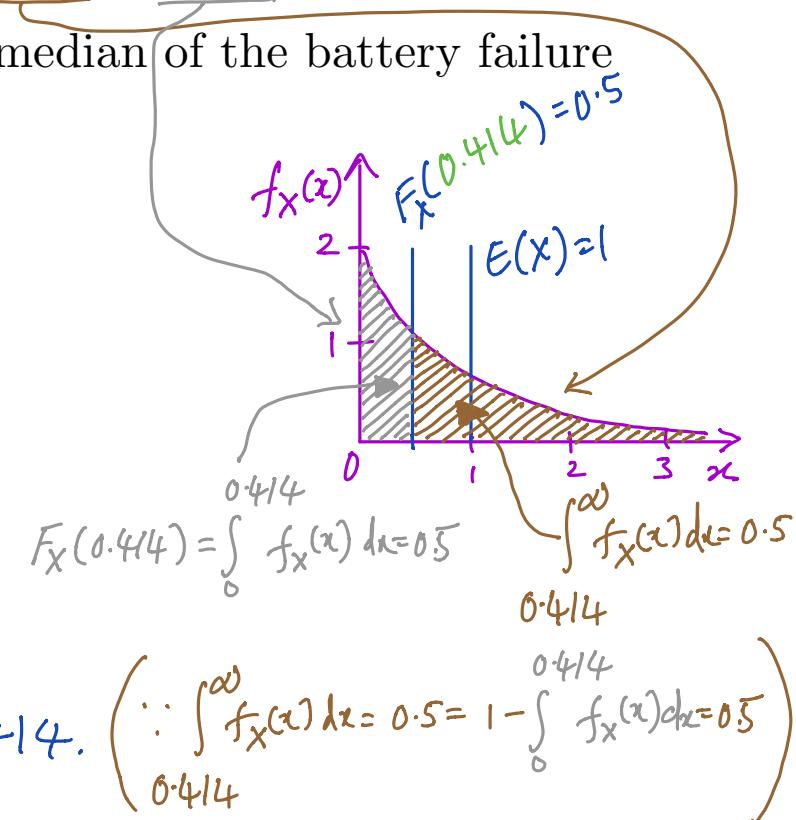
$$\Rightarrow 2[x^2 + 2x + 1 - 1] = x^2 + 2x + 1$$

$$\Rightarrow 2x^2 + 4x = 1 + x^2 + 2x$$

$$\Rightarrow x^2 + 2x + 1 = 1 + 1$$

$$\Rightarrow x+1 = \sqrt{2}$$

$$\Rightarrow x = \sqrt{2} - 1 = 0.414.$$



Continuous random variables

- ▶ Similar to variance for discrete r.v.s, the variance for continuous r.v. X is

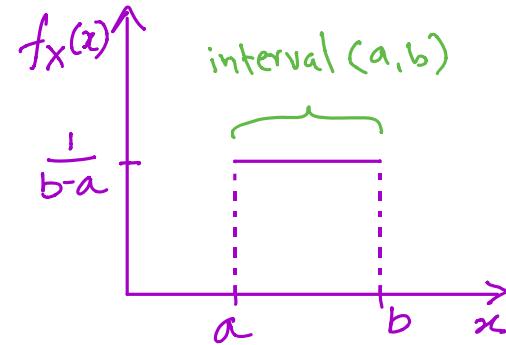
$$\text{Var}(X) = E[(X - E(X))^2] = E(X^2) - E(X)^2.$$

- ▶ Caution: Expectation and Variance are defined only if the integral is absolutely convergent (advance).
- ▶ In this course, we will mainly focus on examples such that expectation and variance are defined.

Continuous random variables

- ▶ Now we will discuss three important continuous distributions: Uniform, Normal (also called Gaussian) and exponential.
- ▶ A continuous r.v. X is said to have the Uniform distribution on the interval (a, b) if its PDF is

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a < x < b; \\ 0, & \text{elsewhere.} \end{cases}$$

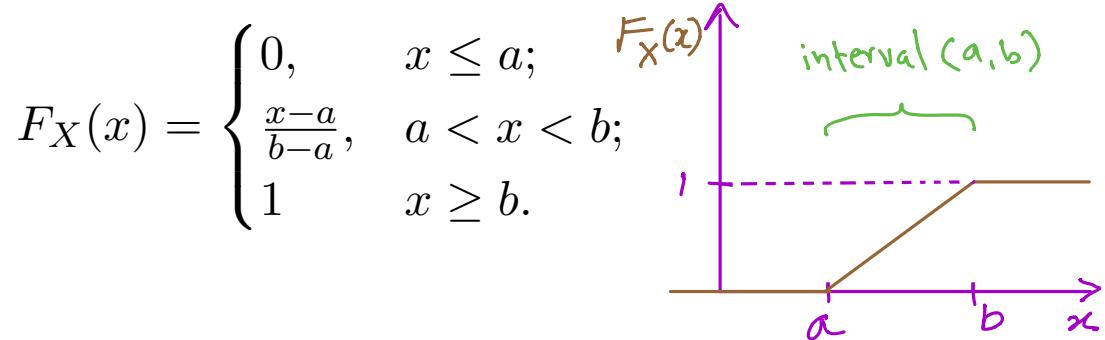


- ▶ Notation: $X \sim \text{Unif}(a, b)$ means that X is uniformly distributed in the interval (a, b) .
- ▶ This is a valid PDF: the area under the curve is just the area of a rectangle with width $b - a$ and height $1/(b - a)$.

$$(1) f_X(x) \geq 0, \quad (2) \int_{-\infty}^{\infty} f_X(x) dx = \int_a^b \frac{1}{b-a} dx = \frac{x}{b-a} \Big|_a^b = \frac{b-a}{b-a} = 1.$$

Continuous random variables

- The CDF is the accumulated area under the PDF:



- For $x \leq a$, $F_X(x) = 0$ since $f_X(x) = 0$ for $x \leq a$.

- For $a < x < b$, $F_X(x) = \int_a^x \frac{1}{b-a} dt = \frac{t}{b-a} \Big|_a^x = \frac{x-a}{b-a}$

- For $x \geq b$, $F_X(x) = \int_a^b f_X(x) dx + \int_b^x \underbrace{f_X(t)}_{0 \text{ for } x \geq b} dt$
 $= \frac{b-a}{b-a} + 0 = 1$

Continuous random variables

- ▶ Example: Find the expectation and variance of $X \sim \text{Unif}(a, b)$.

$$\begin{aligned} E(X) &= \int_a^b x \cdot \frac{1}{b-a} dx \\ &= \frac{x^2}{2(b-a)} \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}. \end{aligned}$$

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

Continuous random variables

$$= \frac{b(a^2 + ab + b^2) - a(a^2 + ab + b^2)}{3(b-a)}$$

$$= \frac{a^2 + ab + b^2}{3}$$

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

$$= \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{2^2}$$

$$= \frac{a^2 + ab + b^2}{3} - \frac{a^2 + 2ab + b^2}{4} = \frac{a^2 - 2ab + b^2}{12}$$
$$= \frac{(a-b)^2}{12}$$

Lecture 20: Continuous Random Variables - Part IV

Satyajit Thakor
IIT Mandi

13 April, 2020

Continuous random variables

- The **normal** or **Gaussian distribution** has a PDF

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

for $-\infty < x < \infty$, depending upon two parameters, the mean and the variance

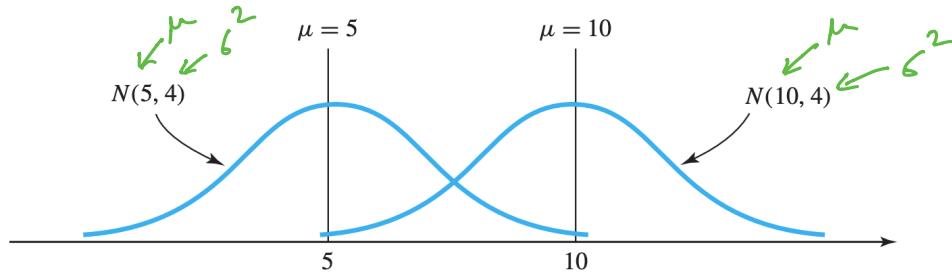
$$E(X) = \mu \text{ and } \text{Var}(X) = \sigma^2$$

of the distribution.

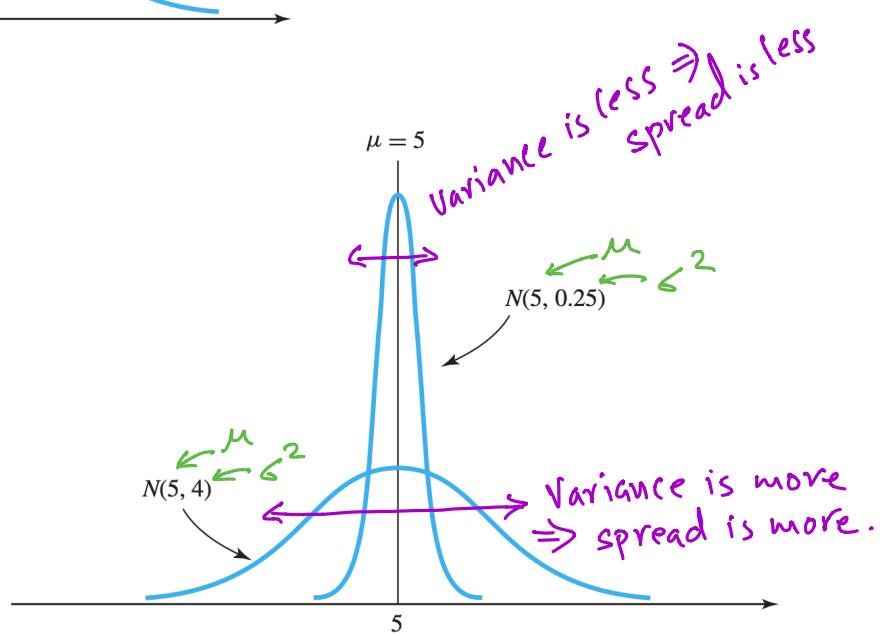
- Notation: $X \sim N(\mu, \sigma^2)$ means that X has a normal distribution with mean μ and variance σ^2 .
- If $X \sim N(0, 1)$ then X is said to have the **standard normal distribution**.

Continuous random variables

- The PDF is a bell-shaped curve that is symmetric about μ :



- Effect of change in variance to the PDF:



Continuous random variables

- The **exponential distribution** is often used to model waiting times (e.g., of a customer in a queue).
- Its PDF is

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0; \\ 0, & x < 0, \end{cases}$$

where $\lambda > 0$ is the parameter for the distribution.

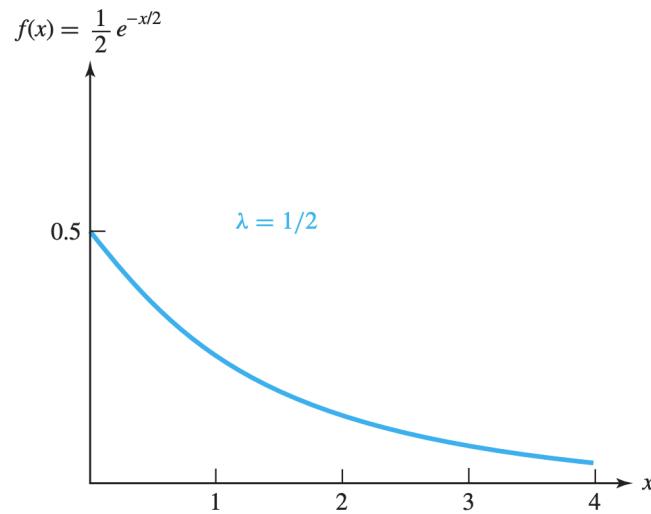
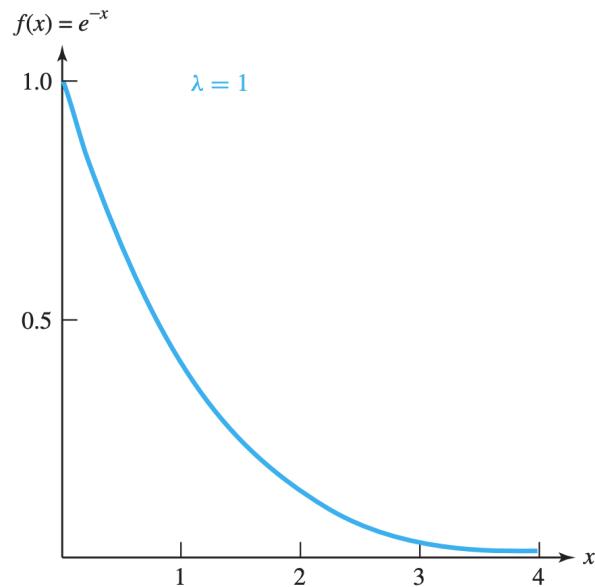
- The CDF is

$$\begin{aligned} F_X(x) &= \int_0^x \lambda e^{-\lambda t} dt, \quad 0 \leq x < \infty \\ &= \left. \frac{\lambda e^{-\lambda t}}{-\lambda} \right|_{t=0}^{t=x} = 1 - e^{-\lambda x} \end{aligned}$$

$$\Rightarrow F_X(x) = \begin{cases} 0 & x < 0, \\ 1 - e^{-\lambda x} & x \geq 0. \end{cases}$$

Continuous random variables

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



Continuous random variables

- Mean of an r.v. with exponential distribution:

$$\begin{aligned} E(X) &= \int_0^\infty t \cdot \lambda e^{-\lambda t} dt \\ &= t \cdot (-e^{-\lambda t}) \Big|_0^\infty - \int_0^\infty -e^{-\lambda t} dt \\ &= 0 - \frac{1}{\lambda} e^{-\lambda t} \Big|_0^\infty \\ &= 0 - \left[0 - \frac{1}{\lambda} \right] \\ &= \frac{1}{\lambda} \end{aligned}$$

Integration by parts:
 $\int u dv = uv - \int v du$

$$\begin{aligned} u &= t &\Rightarrow du = dt \\ v &= -e^{-\lambda t} &\Rightarrow -dv = \lambda e^{-\lambda t} dt \end{aligned}$$

Continuous random variables

- Variance of an r.v. with exponential distribution:

$$E(X^2) = \int_0^\infty t^2 \lambda e^{-\lambda t} dt$$

$\underbrace{t^2}_{u} \underbrace{\lambda e^{-\lambda t}}_{-dv}$

Integration by parts:
 $\int u dv = uv - \int v du$

$$\begin{aligned} &= t^2 \cdot \left(-e^{-\lambda t} \right) \Big|_0^\infty - \int_0^\infty -2t e^{-\lambda t} dt \\ &\quad u = t^2 \Rightarrow du = 2t dt \\ &\quad v = -e^{-\lambda t} \Rightarrow -dv = \lambda e^{-\lambda t} dt \\ &= 0 + \frac{2}{\lambda} \int_0^\infty \lambda t e^{-\lambda t} dt = \frac{2}{\lambda} \cdot \frac{1}{\lambda} = \frac{2}{\lambda^2} \\ &\quad \underbrace{E(X^2)}_{E(X)} \end{aligned}$$

$$\begin{aligned} \Rightarrow \text{Var}(X) &= E(X^2) - E(X)^2 \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}. \end{aligned}$$

Continuous random variables

- Median of an r.v. with exponential distribution:

We need to solve $F_x(x) = 0.5$

$$\Rightarrow 1 - e^{-\lambda x} = 0.5$$

$$\Rightarrow e^{-\lambda x} = 0.5$$

$$\Rightarrow -\lambda x = \ln 0.5$$

$$\Rightarrow -\lambda x = -0.693$$

$$\Rightarrow x = \frac{0.693}{\lambda}$$

$$\Rightarrow x = 0.693 \cdot E(x)$$

Lecture 21: Correlation and Covariance

Satyajit Thakor
IIT Mandi

17 April, 2020

Covariance and correlation

- ▶ When we consider the joint distribution of two random variables, the means, the medians, and the variances of the variables provide useful information about their marginal distributions.
- ▶ However, these values do not provide any information about the dependence between the two variables.
- ▶ The strength of the dependence of two random variables on each other is indicated by their **covariance**, which is defined as

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

Covariance and correlation

► $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$

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

$$= E[XY - XE(Y) - YE(X) + E(X)E(Y)]$$

\downarrow (\because linearity of $E(\cdot)$)

$$= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y)$$

$$= E(XY) - E(X)E(Y)$$

\rightarrow (\because independence
 $\Rightarrow E(XY) = E(X)E(Y)$)

► Independent r.v.s have a covariance of zero.

► The positive covariance indicates a tendency for high values of one random variable to be associated with high values of the other random variable (we shall see this by example).

► Similarly, the negative covariance indicates a tendency for high values of one random variable to be associated with low values of the other random variable.

Covariance and correlation

- The **correlation** between two r.v.s X and Y is defined as

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

- $\text{Corr}(X, Y)$ is also denoted as $\rho(X, Y)$ or $\rho_{X,Y}$. $\Rightarrow (\because \text{independence} \Rightarrow \text{cov}(X, Y) = 0)$
- Independent random variables have a correlation of 0.
- The correlation takes values between -1 and 1. (Why?: proof involves Cauchy–Schwarz inequality - advance topic)
- It is said that:

X and Y are positively correlated if $\text{Corr}(X, Y) > 0$,

X and Y are negatively correlated if $\text{Corr}(X, Y) < 0$, and

X and Y are uncorrelated if $\text{Corr}(X, Y) = 0$.

Covariance and correlation

- ▶ A company services air conditioner units in residences/offices.
- ▶ If the random variable X , is the service time in hours taken at a particular location, and the random variable Y , is the number of air conditioner units at the location, then these two r.v.s can be thought of as jointly distributed.

		X = service time (hrs)				
		1	2	3	4	
$Y = \text{number of air conditioner units}$		1	0.12	0.08	0.07	0.05
		2	0.08	0.15	0.21	0.13
		3	0.01	0.01	0.02	0.07

Covariance and correlation

- ▶ Find the correlation between X and Y . → discussion at the end.
- ▶ Think: Will it be positive or negative? Why?
- ▶ We need to find marginals to compute expectations:

		$X = \text{service time (hrs)}$			
		1	2	3	4
$Y = \text{number of air conditioner units}$	1	0.12	0.08	0.07	0.05
	2	0.08	0.15	0.21	0.13
	3	0.01	0.01	0.02	0.07

$p(X=1) = \sum_{y=1}^3 p(X=1, Y=y)$

$p(Y=1) = \sum_{x=1}^4 p(X=x, Y=1)$

Marginal distribution of X : 0.21, 0.24, 0.30, 0.25

Marginal distribution of Y : 0.32, 0.57, 0.11

Similarly, compute marginals from the joint PMF.

Covariance and correlation

► Now we find expectations:

$$E(X) = \sum_{x=1}^4 x P(X=x) = 1(0.21) + 2(0.24) + 3(0.3) + 4(0.25) \\ = 2.59 \text{ hours}$$

$$E(Y) = \sum_{y=1}^3 y P(Y=y) = 1(0.32) + 2(0.57) + 3(0.11) \\ = 1.79 \text{ units (of AC)}$$

$$E(XY) = \sum_{x=1}^4 \sum_{y=1}^3 xy P(X=x, Y=y) \\ = (1 \cdot 1 \cdot 0.12) + (1 \cdot 2 \cdot 0.08) + \dots + (4 \cdot 3 \cdot 0.07) \\ = 4.86.$$

Covariance and correlation

- Now we find the covariance:

$$\begin{aligned}\text{Cov}(X, Y) &= E(XY) - E(X)E(Y) \\ &= 4.86 - (2.59 \cdot 1.79) \\ &= 0.224.\end{aligned}$$

- The covariance is positive since there is a tendency for locations with a large number of air conditioner units to require relatively long service times. This can also be observed in the joint PMF table.

For example, $P(X=1, Y=3) < P(X=4, Y=3)$

Lecture 22: Continuous Random Variables - Part V

Satyajit Thakor
IIT Mandi

20 April, 2020

Continuous random variables

- Example: A researcher plants 12 seeds whose germination times in days are independent exponential distributions with $\lambda = 0.31$.
- (a) What is the probability that a given seed germinates within five days?

- Recall: If X has exp. distribution then,

$$f_X(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases} \quad \text{and} \quad F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0. \end{cases}$$

- Let X be germination time in days.

- We have X exponentially distributed with $\lambda = 0.31$.

$$\begin{aligned} \text{- Hence, } P(X \leq 5) &= F_X(5) = 1 - e^{-0.31 \cdot 5} \\ &\approx 0.7878 \end{aligned}$$

Continuous random variables

- (b) What are the expectation and variance of the number of seeds germinating within five days?

- We define Y_i as follows:

$$Y_i = \begin{cases} 1 & \text{if } i\text{th seed germinates in 5 days} \\ 0 & \text{" " " does not germinate in 5 days.} \end{cases}$$

- Hence, $P(Y_i=1) = 0.7878$ and $P(Y_i=0) = 1 - P(Y_i=1)$.

- Let $Y = \sum_{i=1}^{12} Y_i$: Y is no. of seeds germinating in 5 days.

- Note that Y is binomially distributed with $n=12, p=0.7878$.

- Hence, $E(Y) = n \cdot p = 12 \cdot 0.7878 = 9.45$.

$$\text{Var}(Y) = np(1-p) = 12 \cdot 0.7878 \cdot 0.2122 = 2.01$$

Continuous random variables

- (c) What is the probability that no more than nine seeds have germinated within five days?

- We need to find $P(Y \leq 9)$.

$$P(Y \leq 9) = \sum_{y=0}^9 P(Y=y)$$

$$= \sum_{y=0}^9 \binom{12}{y} p^y (1-p)^{12-y}$$

$$\approx 0 + 0 + 0 + 0.0051 + 0.0008 + 0.0047 \\ + 0.0202 + 0.0642 + 0.1489 + 0.2457$$

$$= 0.4845.$$

Continuous random variables

- ▶ So far, we have discussed joint and conditional distributions, independence, covariance and correlation for discrete random variables.
- ▶ These parameters can also be extended for continuous random variables.

- ▶ Joint PDF of continuous r.v.s X and Y : the **joint probability density function** is a function $f_{X,Y}(x,y)$ such that $f_{X,Y}(x,y) \geq 0$ and

$$\int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1.$$

- ▶ That is, the PDF is non-negative and the area under the PDF curve is 1.

Continuous random variables

- ▶ Example: a mining company obtains samples of ore from the location and measures their zinc content and their iron content. Suppose that the random variable X is the zinc content of the ore, taking values between 0.5 and 1.5, and that the random variable Y is the iron content of the ore, taking values between 20.0 and 35.0. Furthermore, suppose that their joint PDF is

$$f_{X,Y}(x,y) = \frac{39}{400} - \frac{17(x-1)^2}{50} - \frac{(y-25)^2}{10000}$$

for $0.5 \leq x \leq 1.5$ and $20.0 \leq y \leq 35.0$.

- ▶ Check whether $f_{X,Y}$ is a valid PDF:

- First, check whether $f_{X,Y}(x,y) \geq 0$:

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{39}{400} - \frac{17(x-1)^2}{50} - \frac{(y-25)^2}{10000}, \quad 0.5 \leq x \leq 1.5 \\ &\geq \frac{39}{400} - \frac{17(0.5)^2}{50} - \frac{(10)^2}{10000} = 0.0025 \geq 0. \end{aligned}$$

Continuous random variables

- Now, check whether the area under the curve is 1.

$$\begin{aligned} & \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f_{X,Y}(x,y) dy dx \\ = & \int_{x=0.5}^{1.5} \int_{y=20}^{35} \frac{39}{400} - \frac{17(x-1)^2}{50} - \frac{(y-25)^2}{10000} dy dx \\ = & \int_{x=0.5}^{1.5} \left[\frac{39y}{400} \Big|_{20}^{35} - \frac{17(x-1)^2 \cdot y}{50} \Big|_{20}^{35} - \frac{(y-25)^3}{3 \cdot 16000} \Big|_{20}^{35} \right] dx \\ = & \int_{x=0.5}^{1.5} \left[\frac{39 \cdot 15}{400} - \frac{17 \cdot 15 \cdot (x-1)^2}{50} - \frac{375}{10000} \right] dx \\ = & \frac{39 \cdot 15 \cdot x}{400} \Big|_{0.5}^{1.5} - \frac{17 \cdot 15}{50} \cdot \frac{(x-1)^3}{3} \Big|_{0.5}^{1.5} - \frac{375}{10000} \cdot x \Big|_{0.5}^{1.5} \\ = & \frac{39 \cdot 15}{400} - \frac{17 \cdot 15}{50} \cdot \frac{1}{12} - \frac{375}{10000} = 1. \end{aligned}$$

Continuous random variables

- What is the probability that a randomly chosen sample of ore has a zinc content between 0.8 and 1.0 and an iron content between 25 and 30?

- We want to find $P(0.8 \leq X \leq 1, 25 \leq Y \leq 30)$.

$$P(0.8 \leq X \leq 1, 25 \leq Y \leq 30)$$

$$= \int_{x=0.8}^1 \int_{y=25}^{30} f_{XY}(x, y) dx dy$$

$$= 0.092 \text{ (Homework: verify)}$$

- That is, about 9% of the ore has mineral levels $0.8 \leq X \leq 1, 25 \leq Y \leq 30$.

Lecture 23: Continuous Random Variables - Part VI

Satyajit Thakor
IIT Mandi

21 April, 2020

Continuous random variables

- For two continuous r.v.s, the PDF of the marginal distribution of X is

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

- Example: For the “mining” problem, find the PDF for zinc content of the ore.

- Recall: X : zinc content Y : iron content

$$f_{X,Y}(x, y) = \frac{39}{400} - \frac{17(x-1)^2}{50} - \frac{(y-25)^2}{10000}$$

for $0.5 \leq x \leq 1$, $20 \leq y \leq 35$ and 0 elsewhere.

- We want to find $f_X(x)$.

Continuous random variables

$$\begin{aligned}f_X(x) &= \int_{y=-\infty}^{\infty} f_{X,Y}(x,y) dy \\&= \int_{20}^{35} \left(\frac{39}{400} - \frac{17(x-1)^2}{50} - \frac{(y-25)^2}{10000} \right) dy \\&= \frac{39y}{400} - \frac{17(x-1)^2 y}{50} - \frac{(y-25)^3}{3 \cdot 10000} \Big|_{y=20}^{y=35} \\&= \frac{39 \cdot 15}{400} - \frac{17 \cdot 15(x-1)^2}{50} - \frac{375}{10000} \\&= \frac{14625 - 375}{10000} - \frac{51(x-1)^2}{10} \\&= \frac{57}{40} - \frac{51(x-1)^2}{10}, \quad 0.5 \leq x \leq 1.5 \\&\quad \text{and } 0 \text{ elsewhere.}\end{aligned}$$

Continuous random variables

- If two continuous r.v.s X and Y are jointly distributed, then the conditional distribution of random variable X conditional on the event $Y = y$ has a PDF

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \quad f_Y(y) > 0.$$

- Note that, in the above expression y is fixed and $f_{X|Y=y}$ is a function of the parameter x .
- Example: Suppose that an ore has zinc content of $X = .55$. What is the conditional PDF of the iron content Y given that its zinc content is $X = .55$?

- We want to find $f_{Y|X=.55}(y) = \frac{f_{X,Y}(0.55, y)}{f_X(0.55)}$

Continuous random variables

$$- f_X(0.55) = \frac{57}{40} - \frac{51(0.55-1)^2}{10} = 0.39225$$

- Hence,

$$\begin{aligned}f_{Y|X=0.55}(y) &= \frac{f_{X,Y}(0.55, y)}{0.39225} \\&= \frac{39}{400 \cdot 0.39225} - \frac{17 \cdot (0.55-1)^2}{50 \cdot 0.39225} - \frac{(y-25)^2}{10000 \cdot 0.39225} \\&= 0.073 - \frac{(y-25)^2}{3922.5}\end{aligned}$$

for $20 \leq y \leq 35$ and 0 elsewhere.

Continuous random variables

- Continuous r.v.s X and Y are independent if

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

for all values of x and y .

- If two r.v.s are independent, then

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

- That is, the conditional distributions do not depend upon the value conditioned upon, and they are equal to the marginal distributions.

Continuous random variables

- ▶ Example: Suppose that X and Y have the PDF

$$f_{X,Y}(x,y) = 6xy^2$$

for $0 \leq x \leq 1$ and $0 \leq y \leq 1$ and $f_{X,Y}(x,y) = 0$ elsewhere. Are X and Y independent?

- Let's find f_X , f_Y :

$$f_X(x) = \int_{y=0}^1 6xy^2 dy = \frac{6xy^3}{3} \Big|_{y=0}^1 = 2x$$

$$f_Y(y) = \int_{x=0}^1 6xy^2 dx = \frac{6x^2y^2}{2} \Big|_{x=0}^1 = 3y^2$$

- check: $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) = 6xy^2$.
- Hence, X and Y are independent.

Continuous random variables

- Find the covariance of X and Y with the PDF

$$f_{X,Y}(x,y) = 6xy^2$$

for $0 \leq x \leq 1$ and $0 \leq y \leq 1$ and $f_{X,Y}(x,y) = 0$ elsewhere.

$$\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y)$$

$$E(XY) = \int_{x=0}^1 \int_{y=0}^1 x \cdot y \cdot 6xy^2 dy dx$$

$$= \int_{x=0}^1 \left[\frac{6x^2 y^4}{4} \right]_{y=0}^{y=1} dx$$

$$= \int_{x=0}^1 \frac{6x^2}{4} dx = \left[\frac{6x^3}{12} \right]_{x=0}^{x=1} = \frac{1}{2}$$

Continuous random variables

$$- E(X) = \int_{x=0}^1 x f_X(x) dx = \int_{x=0}^1 x \cdot 2x dx = \frac{2x^3}{3} \Big|_{x=0}^1 = \frac{2}{3}.$$

$$- E(Y) = \int_{y=0}^1 y f_Y(y) dy = \int_{y=0}^1 y \cdot 3y^2 dy = \frac{3y^4}{4} \Big|_{y=0}^{y=1} = \frac{3}{4}.$$

$$\begin{aligned} - \text{Cov}(X, Y) &= E(XY) - E(X) - E(Y) \\ &= \frac{1}{2} - \frac{2}{3} \cdot \frac{3}{4} = 0. \end{aligned}$$

- Since, X and Y are independent, the covariance is zero.

Lecture 24: Inequalities

Satyajit Thakor
IIT Mandi

17 July, 2020

Motivation

- ▶ Let's start with a familiar example: Find $P(|Y| \geq 3)$ if $Y \sim \mathcal{N}(0, 1)$.

standard normal distribution.

$$P(|Y| \geq 3)$$

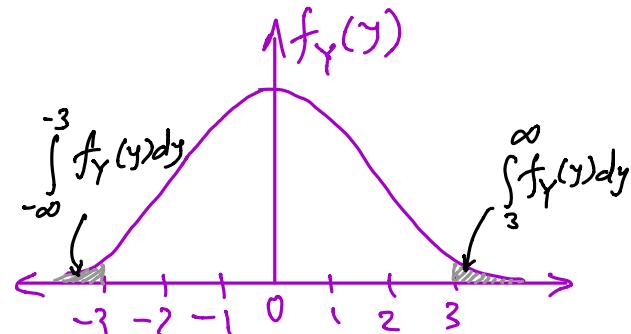
$$= P(Y \leq -3) + P(Y \geq 3)$$

$$= \int_{-\infty}^{-3} f_Y(y) dy + \int_3^{\infty} f_Y(y) dy$$

$$= 0.0013 + 0.0013$$

$$= 0.0026$$

(find using the table for
CDF of $\mathcal{N}(0, 1)$)



Inequalities

- ▶ Many times we like to solve such problems but we may only have limited information.
- ▶ How to solve such problems if we do not know the distribution but only know mean/variance?
- ▶ Markov and Chebyshev inequalities are useful to obtain upper bound on solutions, e.g., $P(|Y| \geq 3)$.
- ▶ Markov inequality: If X is an r.v. that takes only non-negative values, then for any value $a > 0$

$$P(X \geq a) \leq \frac{E(X)}{a}.$$

Proof:

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

Markov inequality

$$= \int_0^\infty x f_X(x) dx \quad (\because X \text{ takes only non-neg. values})$$

$$= \underbrace{\int_0^a x f_X(x) dx}_{\geq 0} + \int_a^\infty x f_X(x) dx$$

$$\geq \int_a^\infty x f_X(x) dx$$

$$\geq \int_a^\infty a f_X(x) dx \quad (\because x \geq a)$$

$$= a \int_a^\infty f_X(x) dx = a P(X \geq a)$$

$$\Rightarrow P(X \geq a) \leq \frac{E(X)}{a}$$

Chebyshev inequality

- If X is an r.v. with mean μ and variance σ^2 , then for any value $a > 0$

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Proof: Note that $|X - \mu| \geq a \iff (X - \mu)^2 \geq a^2$

$$\begin{aligned} \text{Hence, } P(|X - \mu| \geq a) &= P((X - \mu)^2 \geq a^2) \\ &\leq \frac{E[(X - \mu)^2]}{a^2} \quad \left(\because \text{considering } (X - \mu)^2 \text{ as an r.v. and applying Markov ineq.} \right) \\ &= \frac{\sigma^2}{a^2} \end{aligned}$$

Example

- Find an upper bound on $P(|Y| \geq 3)$ if $Y \sim \mathcal{N}(0, 1)$ using (1) Markov inequality (2) Chebyshev inequality.

(1) Using Markov inequality: $P(|Y| \geq 3) \leq \frac{E(|Y|)}{3}$

- Y is $\mathcal{N}(0, 1)$. What is the mean of the r.v. $|Y|$?

$$- E(|Y|) = \int_{-\infty}^{\infty} |y| \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \quad (\because \text{exp. of function of a random variable.})$$

$$= 2 \frac{1}{\sqrt{2\pi}} \int_0^{\infty} y e^{-y^2/2} dy = \sqrt{\frac{2}{\pi}} \int_0^{\infty} y e^{-y^2/2} dy$$

$\underbrace{= 1 \text{ (Homework)}}$

$$- \text{Hence, } P(|Y| \geq 3) \leq \frac{E(|Y|)}{3} = \frac{\sqrt{2/\pi}}{3} = 0.27.$$

Example

(2) Using Chebyshev inequality:

$$P(|Y| \geq 3) \leq \frac{\sigma^2}{3^2}$$

$$= \frac{1}{9}$$

$$= 0.11.$$

- Thus, the chebyshev inequality gives a better upper bound compared to the Markov inequality.
- To use chebyshev ineq. we need μ and σ^2 . and to use markov ineq. we only need μ .

Example

- Suppose that it is known that the number of items produced in a factory during a week is a random variable with mean 50.
 - (a) What can be said about the probability that this week's production will exceed 75?

- Let Y be the no. of produced items in a week.
 - Y is non-negative.
 - Only mean is given.
- \Rightarrow We can use Markov ineq.

$$P(Y \geq 75) \leq \frac{E(Y)}{75}$$

$$= 50/75$$

$$= 2/3$$

Example

- (b) If the variance of a week's production is known to equal 25, then what can be said about the probability that this week's production will be between 40 and 60?

$$\begin{aligned} P(40 < Y < 60) &= P(|Y - 50| < 10) \\ &= 1 - P(|Y - 50| \geq 10) \\ &\geq 1 - \frac{1}{4} \quad \leq \frac{\sigma_Y^2}{10^2} = \frac{25}{100} = \frac{1}{4} \\ &= 3/4 \end{aligned}$$

by Chebyshev ineq

-That is, probability that the production is between 40 and 60 is at least 0.75.

Lecture 25: Weak law of large numbers & Central limit theorem Part I

Satyajit Thakor
IIT Mandi

20 July, 2020

Weak law of large numbers

- ▶ The following statement seems correct by intuition: When the same experiment is repeated many times, the average value obtained over the experiments is close to the expectation of the experiment.
- ▶ Example: How do you know whether a given coin is fair?
- ▶ Another example: consider a Bernoulli random variable with probability of success (i.e., outcome is 1) p . If we conduct n Bernoulli trials then the average of the n outcomes will be very close to p (the expected value of the Bernoulli random variable) for large value of n .
- ▶ This phenomenon is called the weak law of large numbers.

Weak law of large numbers

- **Example:** If we toss a biased coin 400 times with probability of head = .35, it is very likely that approximately 140 times it will be head (i.e., success or outcome 1).



Source: http://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html

Visit this webpage to play with the applet.

Weak law of large numbers

- Let X_1, X_2, \dots be a sequence of independent and identically distributed r.v.s, each having mean $E(X_i) = \mu$. Then, for any $\epsilon > 0$,

$$P\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| > \epsilon\right) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Proof: Note that
$$E\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n} [E(X_1) + \dots + E(X_n)] = \frac{1}{n} \cdot n\mu = \mu$$

$$\text{Var}\left(\frac{X_1 + \dots + X_n}{n}\right) = \left(\frac{1}{n}\right)^2 \cdot \text{Var}(X_1 + \dots + X_n)$$

(∴ linearity of exp., recall Lecture 14)

(∴ $\text{Var}(cX) = c^2 \text{Var}(X)$,
recall Lecture 6)

Weak law of large numbers

$$= \frac{1}{n^2} [Var(X_1) + \dots + Var(X_n)] \quad (\because X_1, \dots, X_n \text{ are independent}$$

recall Lecture 16)

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Now, apply Chebyshhev inequality:

$$P\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| > \epsilon\right) \leq \frac{Var\left(\frac{X_1 + \dots + X_n}{n}\right)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

Note that as $n \rightarrow \infty$, the R.H.S. $\frac{\sigma^2}{n\epsilon^2} \rightarrow 0$. Then, as $n \rightarrow \infty$,

$P\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| > \epsilon\right) \rightarrow 0$ since $P(\cdot)$ is lower bounded by 0 ($\because P(\cdot)$ is a non-neg. function) and upper bounded by $\frac{\sigma^2}{n\epsilon^2}$.

A result on normal distributions

- ▶ Before we study the central limit theorem, let's look at the relation between a normal distribution $\mathcal{N}(\mu, \sigma^2)$ and the standard normal distribution $\mathcal{N}(0, 1)$.
- ▶ Recall that if we want to solve problems on the standard normal distribution then we refer to its CDF table.
- ▶ What to do if we want to solve a problem on a normal distribution $\mathcal{N}(\mu, \sigma^2)$?
- ▶ Relation: If $Z \sim \mathcal{N}(0, 1)$, then $X = \mu + \sigma Z$ have the normal distribution $\mathcal{N}(\mu, \sigma^2)$. (proof: next slide) 
- ▶ An application of this relation: note that $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$. Hence, the CDF of X can be found from the CDF of the standard normally distributed Z .

A result on normal distributions

- Relation: If $Z \sim \mathcal{N}(0, 1)$, then $X = \mu + \sigma Z$ have the normal distribution $\mathcal{N}(\mu, \sigma^2)$.

Partial Proof: We will only prove that $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$.

$$\begin{aligned} E(X) &= E(\mu + \sigma Z) \\ &= E(\mu) + \sigma E(Z) \\ &= \underbrace{\mu}_{\text{linearity of exp.}} + \sigma \cdot 0 \\ &= \mu \end{aligned}$$

vercall lecture 14

$$\begin{aligned} \text{Var}(X) &= \text{Var}(\mu + \sigma Z) \\ &= \text{Var}(\sigma Z) \quad \because \text{Var}(x+c) = \text{Var}(x) \\ &= \sigma^2 \underbrace{\text{Var}(Z)}_1 \quad \because \text{Var}(cX) = c^2 \text{Var}(X) \\ &= \sigma^2 \end{aligned}$$

vercall lecture 16

A result on normal distributions

μ σ^2

- Example (application of the relation): Let $X \sim N(-1, 4)$. What is $P(|X| < 3)$?

normal but not standard normal.

- We know that if $X \sim N(\mu, \sigma^2)$ then $X = \mu + \sigma Z$
where $Z \sim N(0, 1)$. $\Rightarrow Z = \frac{X-\mu}{\sigma} \Rightarrow Z = \frac{X+1}{2}$.

$$\begin{aligned} \text{Now, } P(|X| < 3) &= P(-3 < X < 3) \\ &= P\left(\frac{-3+1}{2} < \frac{X+1}{2} < \frac{3+1}{2}\right) \\ &= P(-1 < Z < 2) \quad \left(z \sim N(0, 1)\right) \\ &= F_Z(2) - F_Z(-1) \quad \left(\Rightarrow \text{use the CDF table}\right) \\ &= .9772 - 0.1587 = 0.8185 \end{aligned}$$

Lecture 26: Weak law of large numbers & Central limit theorem Part II

Satyajit Thakor
IIT Mandi

22 July, 2020

Example

- ▶ Example (application of Chebyshev inequality): The time taken to serve a customer at a fast-food restaurant has a mean of 75.0 seconds and a standard deviation of 7.3 seconds. Use Chebyshev inequality to calculate the time interval that has 75% probability of containing a particular service time.

- Chebyshev inequality: $P(|X-\mu| \geq a) \leq \frac{6^2}{a^2}$.
- Let X be the time taken to serve a customer.
- Find a time interval that has 75%. Prob. of containing a particular service time.
- Note that, $P(|X-\mu| \geq a) = 1 - P(|X-\mu| < a)$
 $\Rightarrow 1 - P(|X-\mu| < a) \leq \frac{6^2}{a^2}$
 $\Rightarrow P(|X-\mu| < a) \geq 1 - \frac{6^2}{a^2}$

Example

- To find the desired interval, let

$$1 - \frac{\zeta^2}{a^2} = 0.75 \Rightarrow a^2 = \frac{(7.3)^2}{.25} = 213.16$$
$$\Rightarrow a = 14.6$$

- Hence, $P(|X - \mu| < 14.6) \geq 0.75$

$$\Rightarrow P(-14.6 < X - 75 < 14.6) \geq 0.75$$

$$\Rightarrow P(60.4 < X < 89.6) \geq 0.75.$$

- Thus, for X in the interval $(60.4, 89.6)$ the probability is at least 75%. (by chebyshev inequality)

Central limit theorem

- ▶ Consider a sequence X_1, \dots, X_n of independent identically distributed random variables. Suppose that $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$ for $i = 1, \dots, n$.
- ▶ If we define

average r.v. : $\bar{X} = \frac{X_1 + \dots + X_n}{n}$

then $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

} proved in Lecture 25.

- ▶ The central limit theorem states that regardless of the actual distribution of the individual random variables X_i , the distribution of their average \bar{X} is closely approximated by a $\mathcal{N}(\mu, \sigma^2/n)$ distribution.
- ▶ That is, average of a set of independent identically distributed random variables is always approximately normally distributed.
- ▶ The accuracy of the approximation improves as n increases.

Central limit theorem (CLT)



- If X_1, \dots, X_n is a sequence of independent identically distributed random variables with a mean μ and a variance σ^2 , then the distribution of their average \bar{X} can be approximated by

$$\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

- This is equivalent to: If X_1, \dots, X_n is a sequence of independent identically distributed random variables with a mean μ and a variance σ^2 , then the distribution of the sum $X_1 + \dots + X_n$ can be approximated by

$$\mathcal{N}(n\mu, n\sigma^2).$$

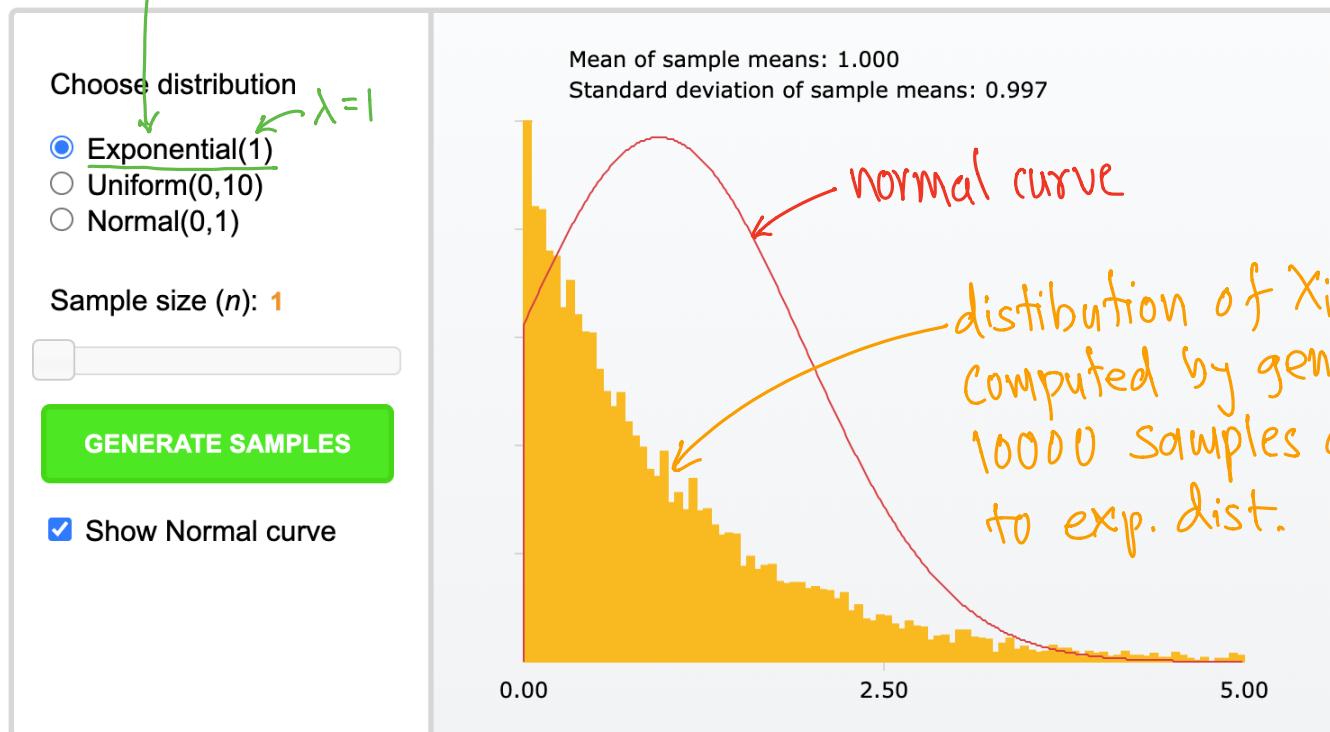
- The proof is beyond the scope of this course. We will study the statement with examples.

theorem statement

Central limit theorem

$$f_X(x) = \lambda e^{-\lambda x}, x \geq 0. \quad \mu_X = \frac{1}{\lambda}, \quad \sigma^2 = \frac{1}{\lambda^2}.$$

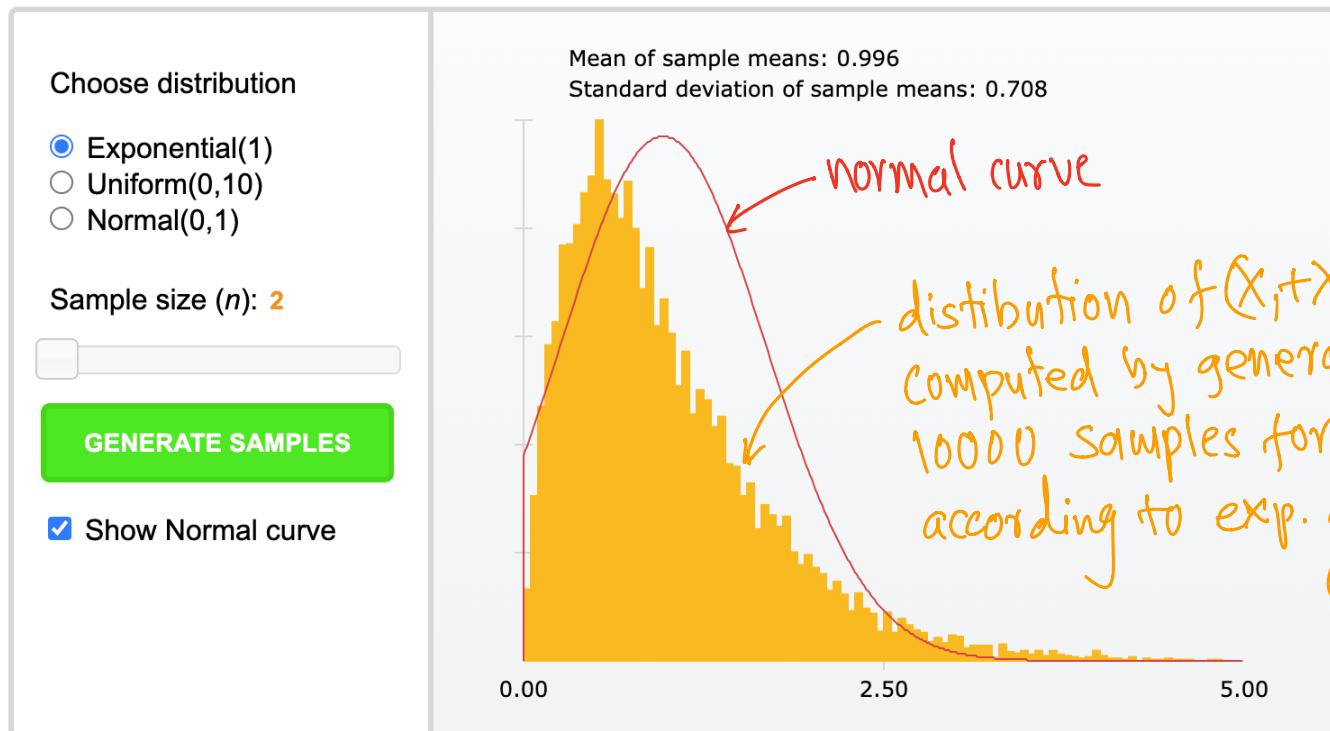
- Numerical exponential distribution: X_i



Source: https://digitalfirst.bfwpub.com/stats_applet/stats_applet_3_cltmean.html

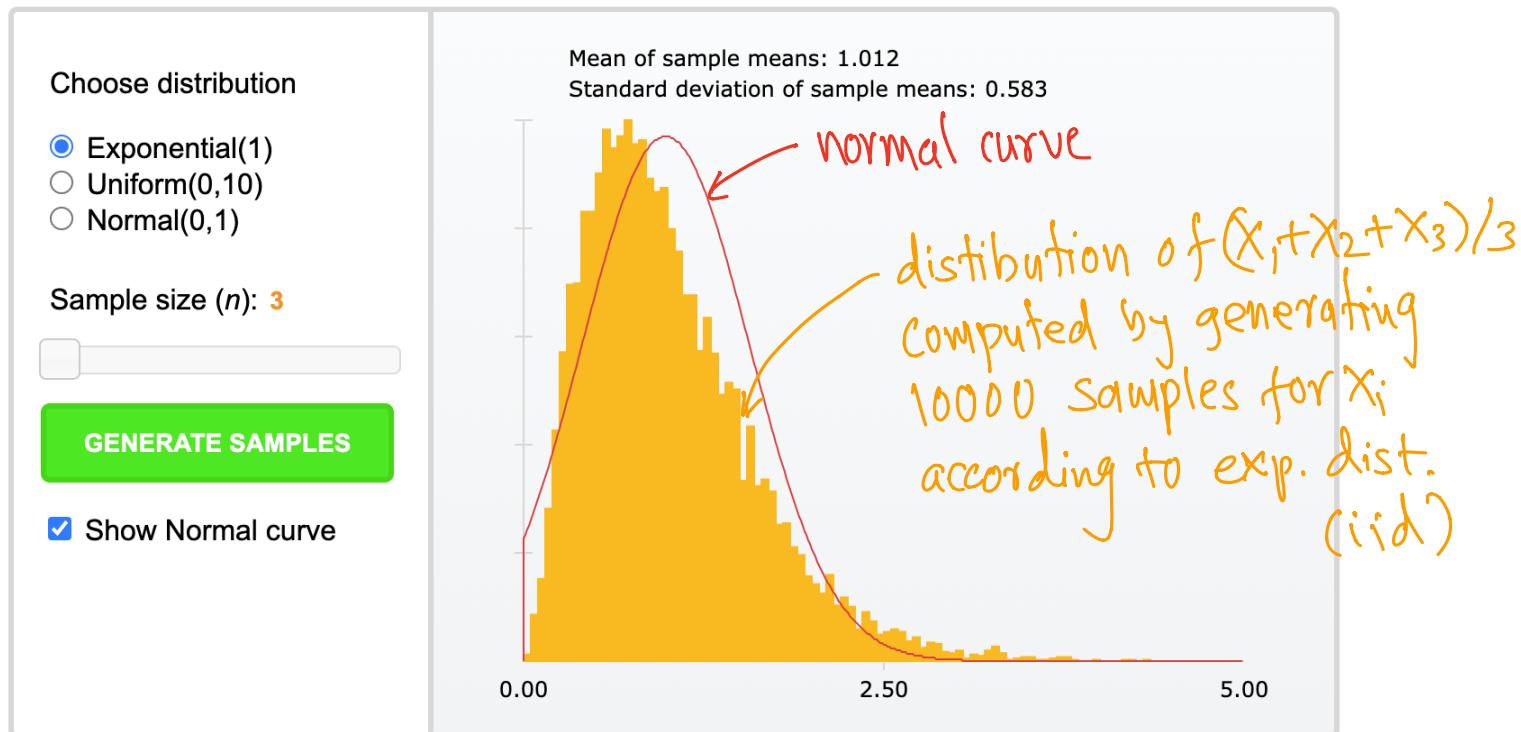
Central limit theorem

- Numerical exponential distribution: $(X_1 + X_2)/2$



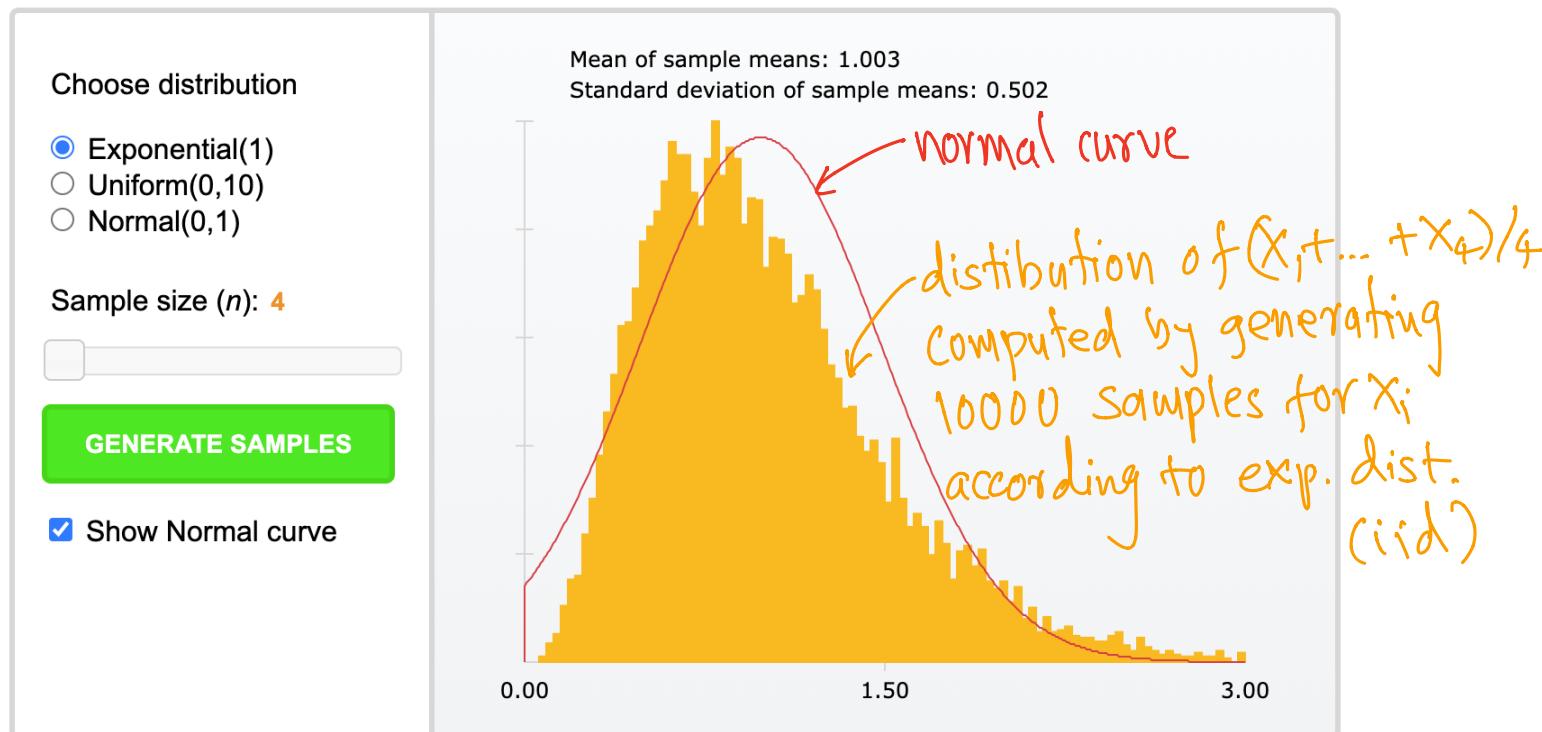
Central limit theorem

- Numerical exponential distribution: $(X_1 + X_2 + X_3)/3$



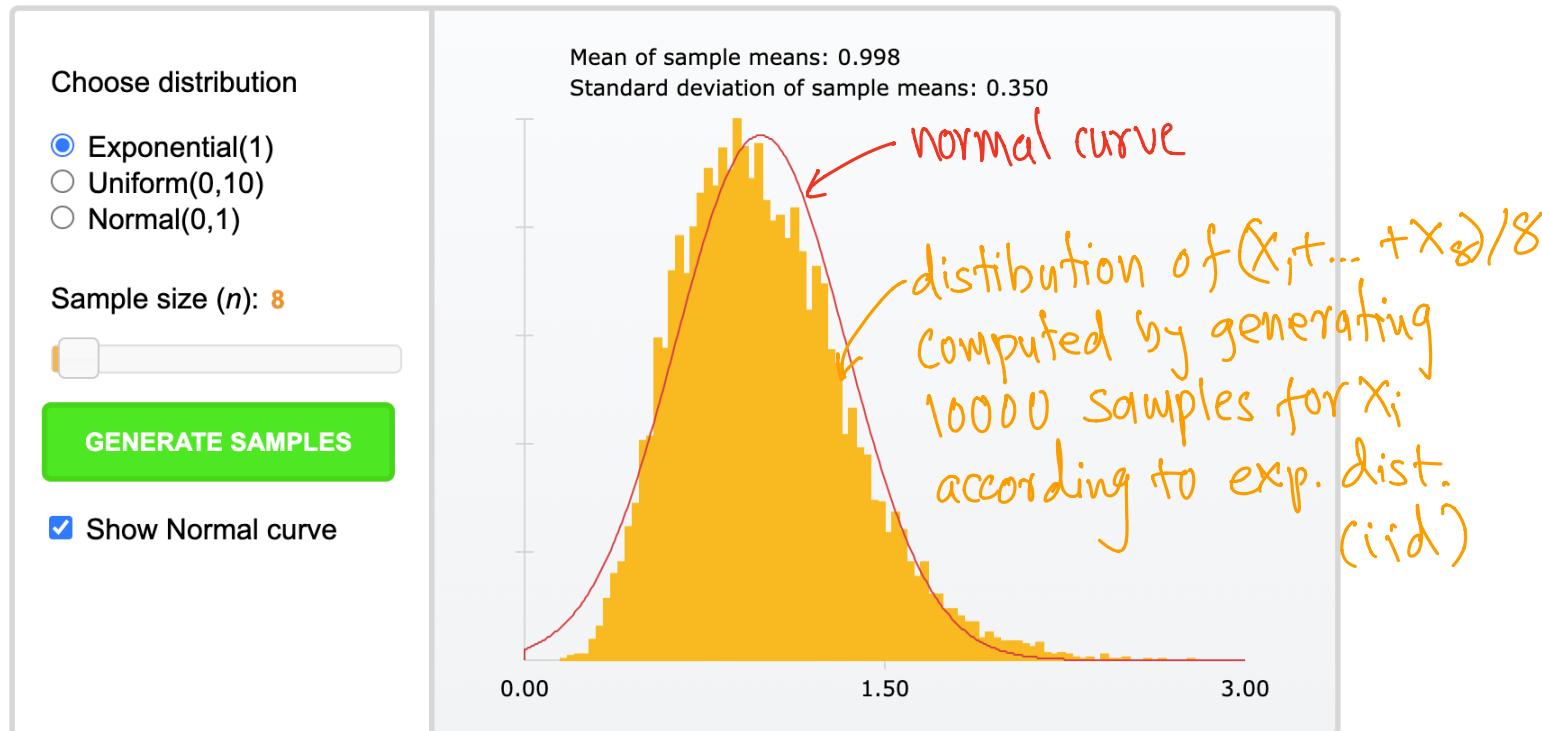
Central limit theorem

- Numerical exponential distribution: $(X_1 + \dots + X_4)/4$



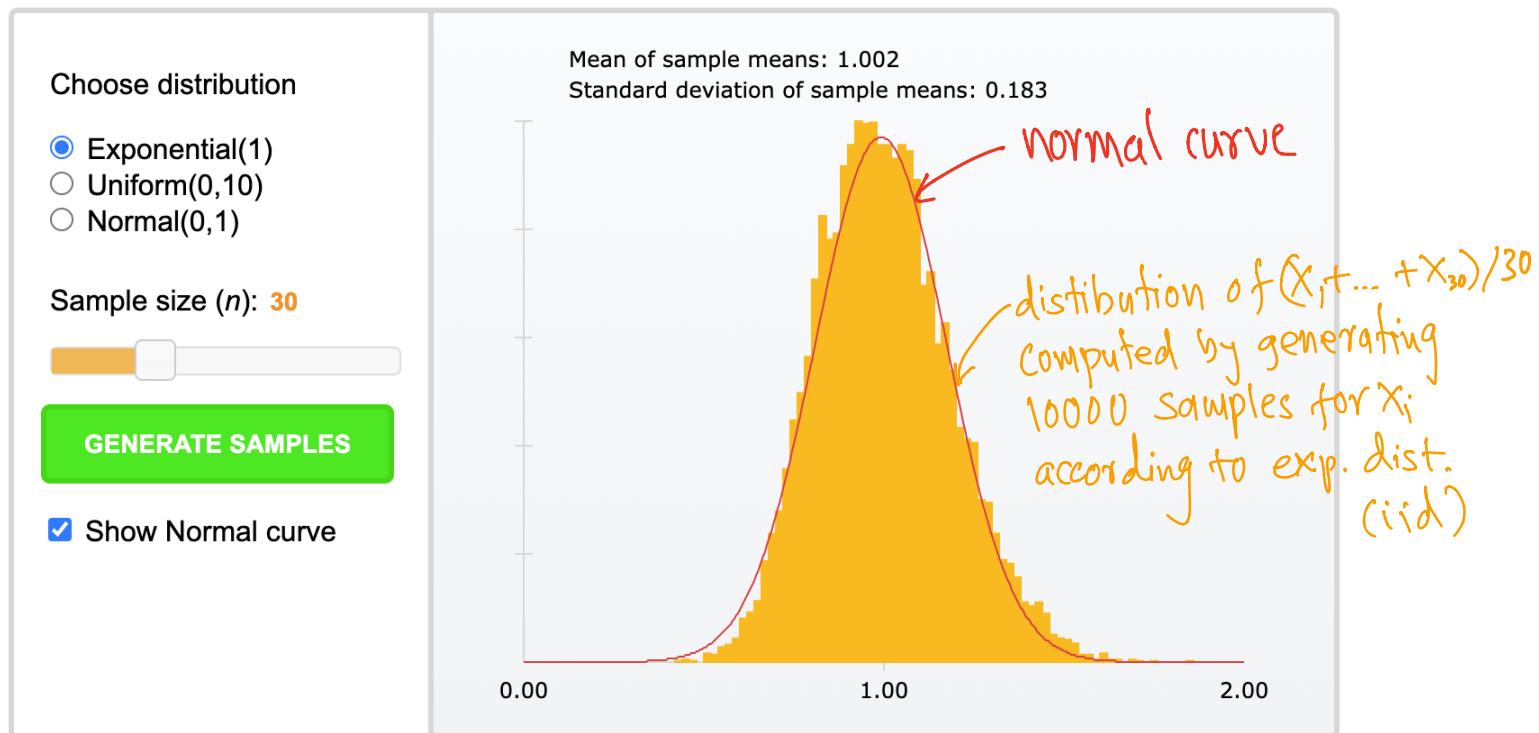
Central limit theorem

- Numerical exponential distribution: $(X_1 + \dots + X_8)/8$



Central limit theorem

- Numerical exponential distribution: $(X_1 + \dots + X_{30})/30$



Example: CLT

- ▶ Recall Problem 2, Assignment 6: Milk containers have label printed “2 liters”. But, the PDF of the amount of milk deposited in a milk container by a dairy factory is

$$f_X(x) = \begin{cases} 40.976 - 16x - 30e^{-x}, & 1.95 \leq x \leq 2.20; \\ 0, & \text{elsewhere.} \end{cases}$$

- (a) Is f_X a valid PDF?
 - (b) What is the probability that a container produced by the dairy factory is underweight?
- ▶ Recall that the probability of “a container produced by the dairy factory is underweight” is .261 (solution of 2(b)).

Example: CLT

- ▶ Example: What is the distribution of the number of underweight containers X in a box of 20 containers? Find the
 - (a) exact and
 - (b) approximate (using CLT) value of the probability "a box contains no more than three underweight containers".

A container is underweight, with prob. .261.

denote as r.v. W : this is a Bernoulli r.v. with $p = 0.261$.
 $W=1$ (underweight) with prob. P
 $W=0$ (not underweight) with prob. $1-p$.

No. of underweight containers in a box of 20.
denote as r.v. X : this is a Binomial r.v. $B(20, \frac{.261}{n})$

Recall: sum of iid Bernoulli r.v.s is Binomial.

Example: CLT

$$(a) P(X \leq 3) = \sum_{k=0}^3 \binom{20}{k} (.261)^k (1-.261)^{20-k} = \underbrace{0.1934}_{\text{exact}}$$

(b) Note that a Binomial r.v. with the distribution $B(n, p)$ is a sum of n i.i.d Bernoulli r.v.s. Hence, by the CLT, we can approximate :

$$X \sim B(n, p) \approx N(n\mu_W, n\sigma_W^2)$$

$$\Rightarrow X \sim B(20, .261) \approx N(20 \times .261, 20 \times .261 \times (1-.261)) \\ \approx N(5.22, 3.86) \sim Y = \mu + \sigma Z$$

$$\Rightarrow P(X \leq 3) = P(X \leq 3.5) \approx P(Y \leq 3.5) \\ = P(\mu + \sigma Z \leq 3.5) \\ = P(5.22 + \sqrt{3.86} Z \leq 3.5) \\ = P(Z \leq -0.8754) = 0.1922$$

Very close to the exact value.

Lecture 27: Introduction to Statistics

Satyajit Thakor
IIT Mandi

28 July, 2020

Statistical inference

- ▶ Till now we studied probability theory, i.e., random variables through their distributions.
- ▶ In practice, often an experimenter has no knowledge of distribution.
- ▶ A task of the experimenter is to find out as much information as possible about the distribution.
- ▶ This is done through experimentation and the collection of a data set relating to the random variable.
- ▶ Statistical inference is the science of deducing properties of an underlying probability distribution from such a data set.

Population and sample

- ▶ A population consists of all possible observations available from a particular probability distribution.
- ▶ Example: let X be the weight of a milk container produced at a particular dairy. Then, weight of each container is the population.
- ▶ A sample is a particular subset of the population that an experimenter measures and it is used to investigate the unknown distribution.
- ▶ Example: a sample or data set for “the weight of a milk container” is obtained by weighing the contents of n containers.
- ▶ A random sample is one in which the elements of the sample are chosen at random from the population, and this procedure is often used to ensure that the sample is representative of the population.
- ▶ Example: if we choose the first n containers produced then this does not provide random sample.

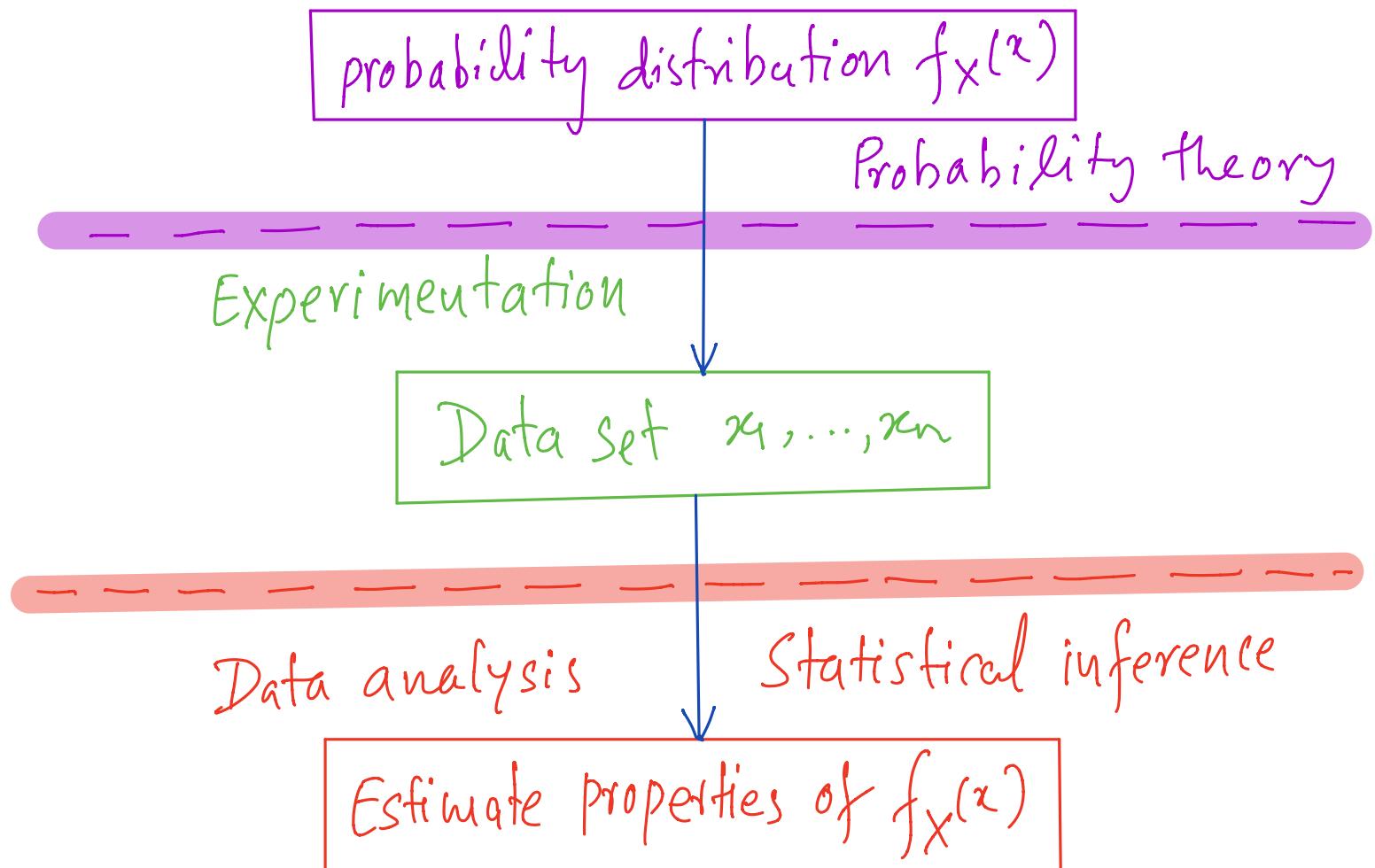
Population and sample

- ▶ The PDF $f_X(x)$ provides complete information about the probabilistic properties of the random variable X and is unknown to the experimenter.
- ▶ The experimenter proceeds by obtaining a sample of observations of the random variable X , which may be written

$x_1, x_2, \dots, x_n.$ sample / data set

- ▶ An appropriate analysis of the data gives the experimenter some information about $f_X(x)$.

Probability theory & statistical inference



Sample mean

- ▶ The sample mean of a data set is simply the arithmetic average of the data observations.
- ▶ That is, if a data set consists of the n observations x_1, \dots, x_n , then the sample mean is

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- ▶ The sample mean \bar{x} can be thought of as being an estimate of the expectation of the unknown underlying probability distribution of the observations in the data set.

Sample variance

- ▶ The sample variance of a set of data observations x_1, \dots, x_n is defined as

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}.$$

- ▶ The sample standard deviation is s .
- ▶ Why the denominator of the formula for s^2 is chosen as $n - 1$ and not n ? -We will see this in the next lecture.

Parameter and statistics

- ▶ A **parameter** is a property of a population or a probability distribution.
- ▶ For example, the PDF of a population of r.v. X is $f_X(x)$ and μ_X is a parameter.
- ▶ A **statistic** is a property of a sample from the population.
- ▶ For example, suppose that a sample of size n is collected of observations from a particular probability distribution $f_X(x)$. The data values recorded, x_1, \dots, x_n , are the observed values of a set of n random variables X_1, \dots, X_n , and each has the probability distribution $f_X(x)$.

Parameter and statistics

- ▶ In general, a statistic is any function $g(X_1, \dots, X_n)$ of these random variables.
- ▶ For example, the sample mean

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}$$

and the sample variance

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \text{ are } \underline{\text{statistics.}}$$

- ▶ For a given data set x_1, \dots, x_n these statistics take the observed values

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \text{ and } s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}.$$

Lecture 28: Estimation

Satyajit Thakor
IIT Mandi

31 July, 2020

Recall

- ▶ A population is the set of all possible observations available from a particular probability distribution.
- ▶ A sample is a particular subset of the population.

e.g., mean, variance of unknown PDF.

- ▶ A parameter is a property of a population or a probability distribution.
- ▶ A statistic is a property of a sample from the population.

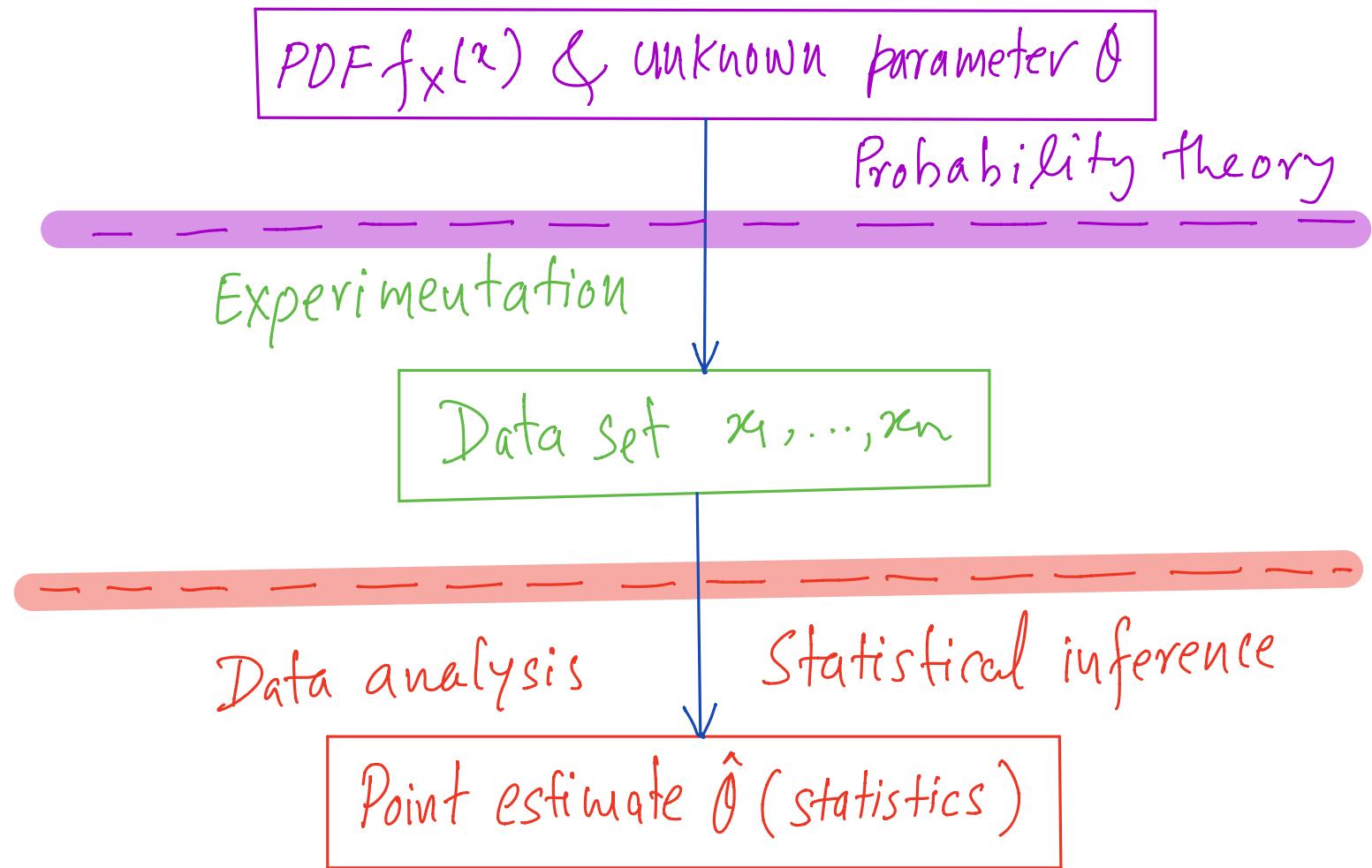
e.g., sample mean, sample variance

Point estimation

- ▶ Estimation is a procedure by which the information contained within a sample is used to investigate properties of the population (parameters) from which the sample is drawn.
 - ▶ A point estimate of an unknown parameter θ is a statistic $\hat{\theta}$ that represents a “best guess” at the value of θ .
 - ▶ There may be more than one sensible point estimate of a parameter.
 - ▶ Example: point estimate of the mean μ_X of a probability distribution $f_X(x)$ is the sample mean \bar{x} of data observations obtained from the probability distribution. In this case, $\hat{\mu}_X = \bar{x}$.
 - ▶ Similarly, two point estimates of the variance σ_X^2 are

Similarly, two point estimates of the variance of X are

Point estimation



Point estimation

- ▶ Recall that $\hat{\theta}$ is a point estimate of a parameter θ which is a function $g(x_1, \dots, x_n)$. Similarly, $\hat{\Theta}$ is a point estimate which is a function $g(X_1, \dots, X_n)$. Note that $\hat{\theta}$ is a number and $\hat{\Theta}$ is an r.v.
- ▶ In general, when there is more than one obvious point estimate for a parameter (e.g., estimates of variance in Slide 3), the following criteria can be used to find desirable point estimate:
- ▶ A point estimate $\hat{\Theta}$ for a parameter θ is called **unbiased** if

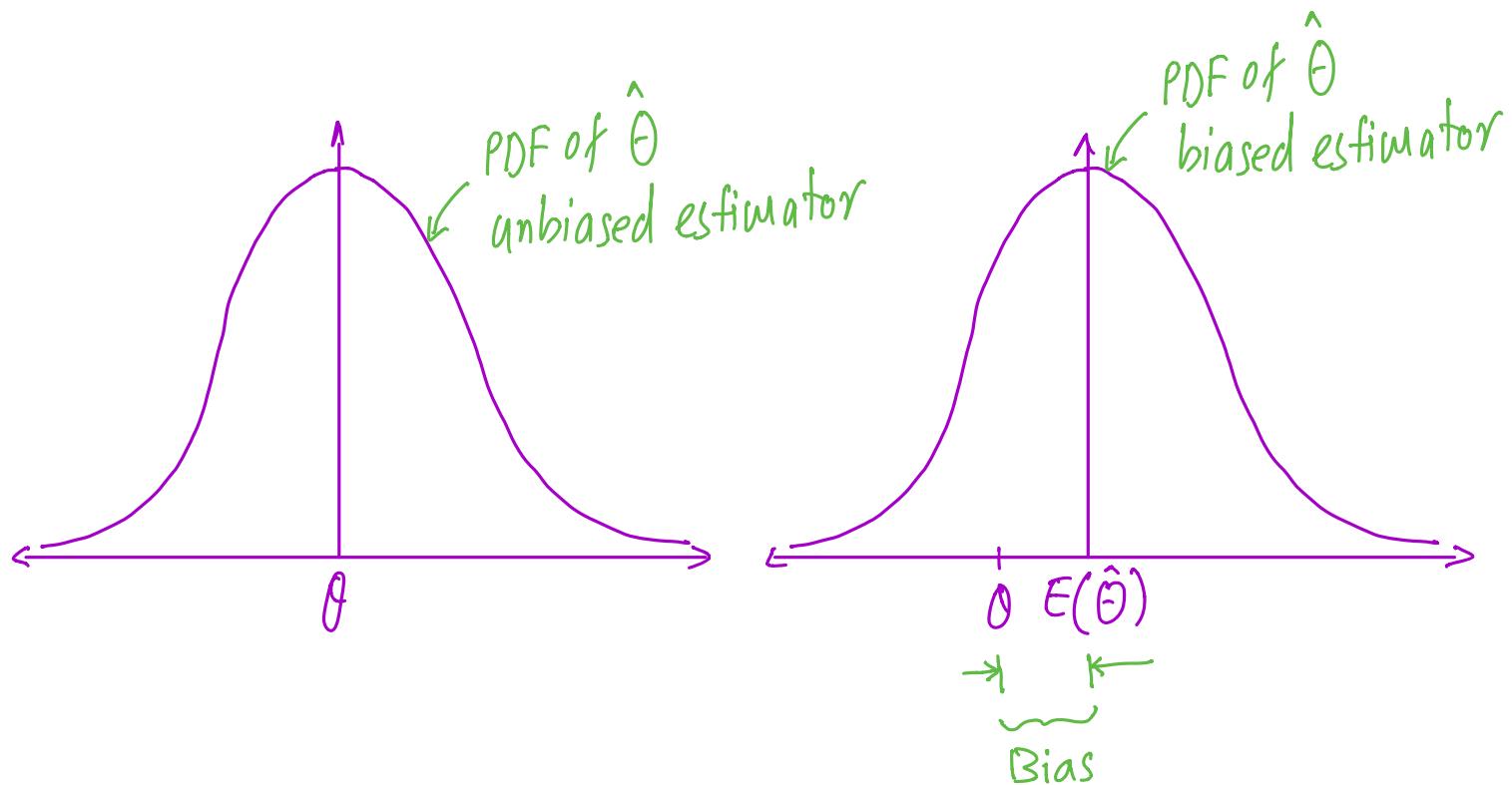
$$E(\hat{\Theta}) = \theta$$

- ▶ Unbiasedness is a nice property for a point estimate to possess. If a point estimate is not unbiased, then its **bias** can be defined as

$$\text{bias} = E(\hat{\Theta}) - \theta.$$

Point estimation

- Unbiased estimator: $E(\hat{\theta}) = \theta$
- Biased estimator: $E(\hat{\theta}) \neq \theta$, bias = $E(\hat{\theta}) - \theta$



Point estimation

- Example: If X_1, \dots, X_n is a sample of observations from a probability distribution with a mean μ_X , then show that the sample mean $\hat{\Theta} = \hat{\mu}_X = \bar{X}$ is an unbiased point estimate of the population mean $\theta = \mu_X$.

Proof: Note that $E(X_i) = \mu_X$ for all $i=1, \dots, n$.

$$\begin{aligned}\text{Hence, } E(\hat{\mu}_X) &= E(\bar{X}) \\ &= E\left(\frac{X_1 + \dots + X_n}{n}\right) \\ &= \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{n\mu_X}{n} = \mu_X.\end{aligned}$$

$E(\hat{\mu}_X) = \mu \Rightarrow \hat{\mu}_X$ is unbiased p.e. of μ_X .

Point estimation

- Example: If X_1, \dots, X_n is a sample of observations from a probability distribution with a variance σ_X^2 , then show that the sample variance

$$\hat{\Theta} = \hat{\sigma}_X^2 = S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

is an unbiased point estimate of the population variance $\theta = \sigma_X^2$.

Proof: $E(S^2) = \frac{1}{n-1} E\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)$

You may skip
the proof.
(Won't be asked
in the exams)

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

Point estimation

$$= \frac{1}{n-1} E \left(\sum_{i=1}^n (x_i - \mu_x)^2 - 2(\bar{x} - \mu_x) \underbrace{\sum_{i=1}^n (x_i - \mu_x)}_{= n(\bar{x} - \mu_x)} + n(\bar{x} - \mu_x)^2 \right)$$

$$= \frac{1}{n-1} E \left(\sum_{i=1}^n (x_i - \mu_x)^2 - n(\bar{x} - \mu_x)^2 \right)$$

$$= \frac{1}{n-1} \left(\underbrace{\sum_{i=1}^n E[(x_i - \mu_x)^2]}_{\text{Var}(x_i) = \sigma_x^2} - n E[(\bar{x} - \mu_x)^2] \underbrace{\text{Var}(\bar{x})}_{\text{Var}(\bar{x}) = \sigma_x^2/n} \right)$$

$$= \frac{1}{n-1} \left(n\sigma_x^2 - n\sigma_x^2/n \right) = \sigma_x^2$$

$\Rightarrow s^2$ is an unbiased estimator of σ_x^2 .

Point estimation

- Homework (Assignment 9 problem): If X_1, \dots, X_n is a sample of observations from a probability distribution with a variance σ_X^2 , then show that the sample variance

Hint: use the fact that

$$\hat{\Theta} = \hat{\sigma}_X^2 = S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

$$\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$$

has the bias of

$$-\frac{\sigma_X^2}{n}$$

is an unbiased estimator.

for point estimate of the population variance $\theta = \sigma_X^2$.

- Hence, we choose the denominator $n - 1$ to make sure that the estimator is unbiased.

(Recall : Slide 7, Lecture 27)

Lecture 29: Estimation - Part II

Satyajit Thakor
IIT Mandi

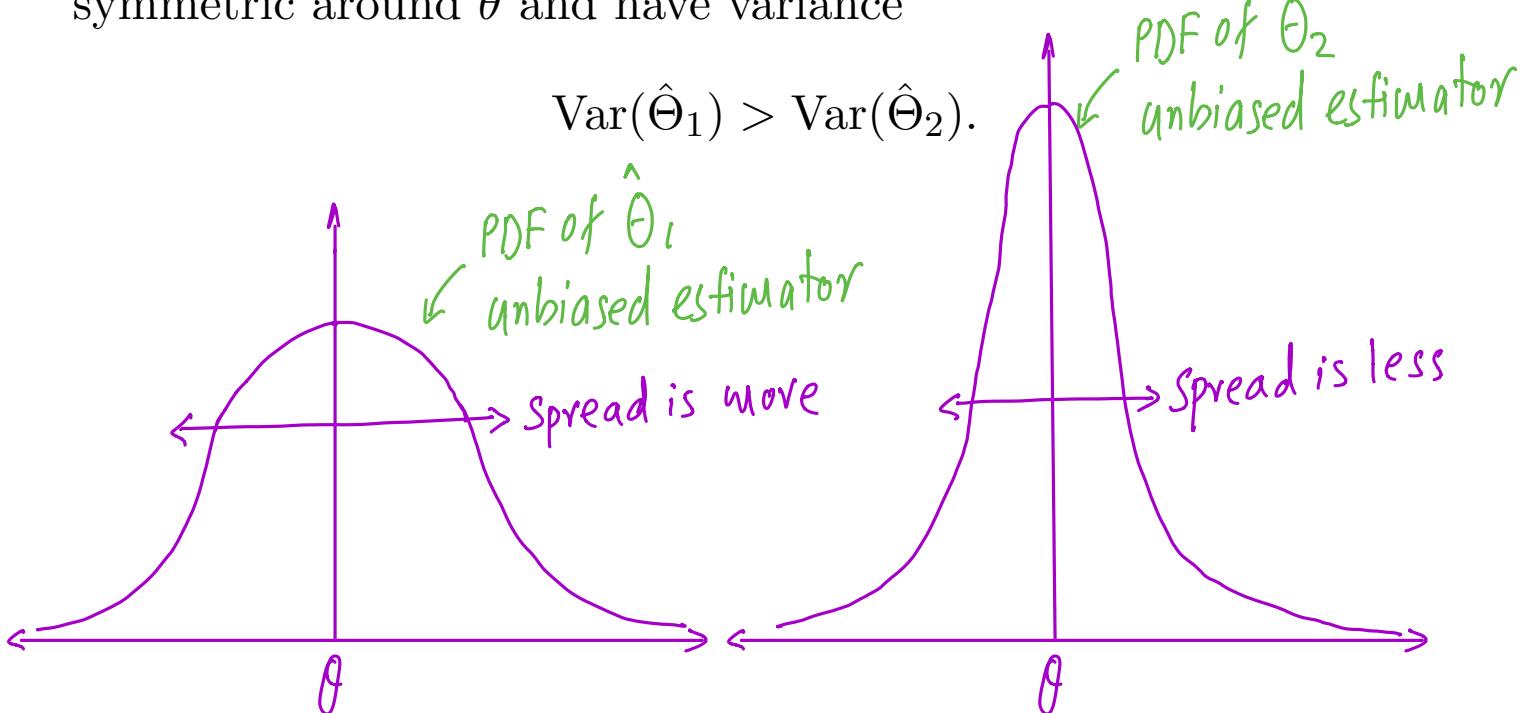
4 August, 2020

Point estimation

$$E(\hat{\theta}) = \theta$$

- Recall that an unbiased estimator is more desirable in practice.
- If we have two or more unbiased estimators to choose from, which is better to choose?
- Suppose that two point estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased and symmetric around θ and have variance

$$\text{Var}(\hat{\theta}_1) > \text{Var}(\hat{\theta}_2).$$

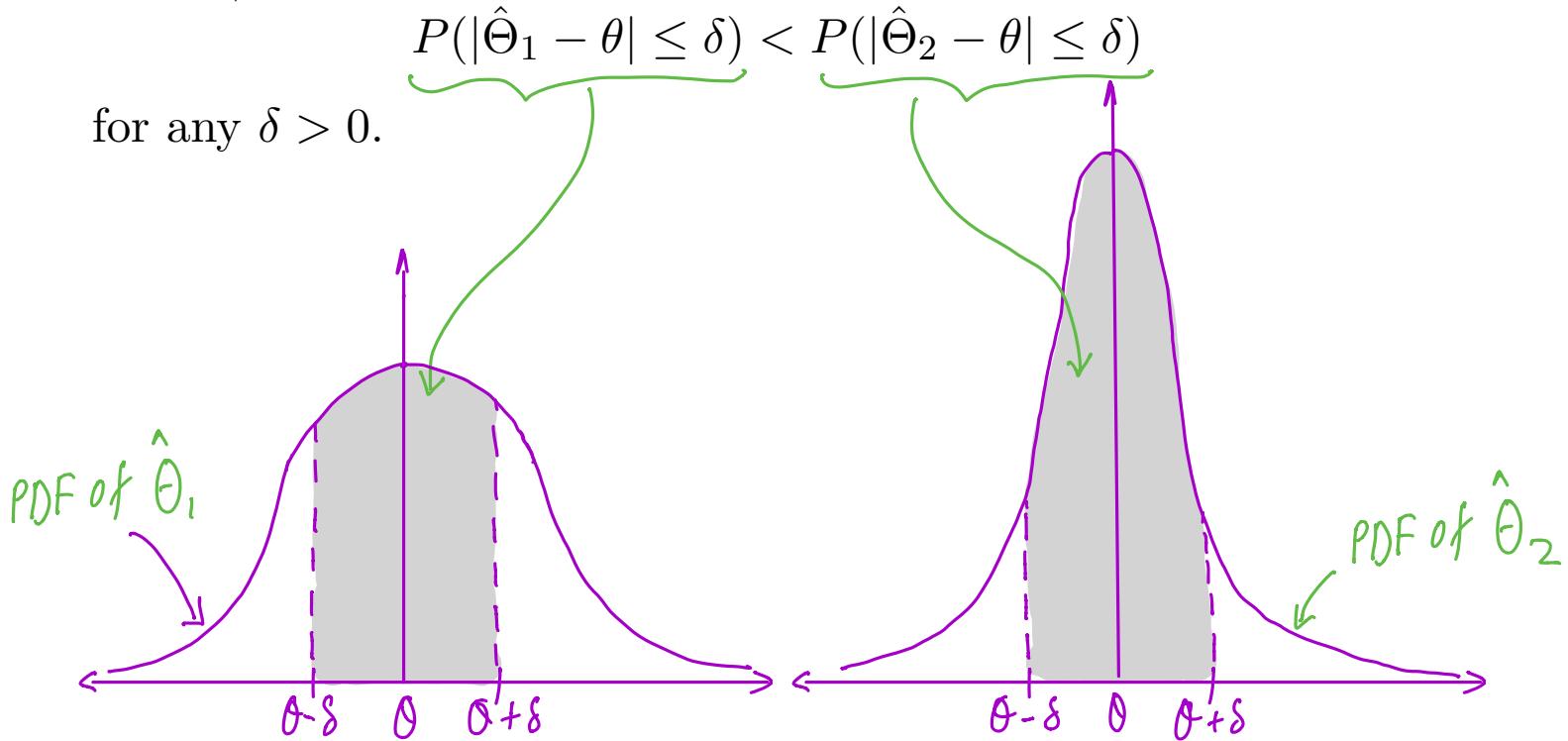


Point estimation

- ▶ Which is a better point estimate? Answer: $\hat{\Theta}_2$
- ▶ $\hat{\Theta}_2$ is better in the sense that it is likely to provide an estimate closer to the true value θ than the estimate provided by $\hat{\Theta}_1$, that is,

$$P(|\hat{\Theta}_1 - \theta| \leq \delta) < P(|\hat{\Theta}_2 - \theta| \leq \delta)$$

for any $\delta > 0$.



Maximum likelihood estimation

- ▶ Till now we discussed simple ways to define an estimate, e.g., sample mean, sample variance.
- ▶ Now we will study one of the most popular (and scientific) method for parameter estimation.
- ▶ Some times, it happens that you know the family distribution of the random variable, e.g., Bernoulli, binomial, normal, etc.
- ▶ But you do not know the crucial parameter. For example:
 - ▶ In $\text{Bern}(p)$, p is not known.
 - ▶ In $\text{Binom}(n, p)$, p is not known.
 - ▶ In $\text{Pois}(\lambda)$, λ is not known.
 - ▶ In $\mathcal{N}(\mu, \sigma^2)$, μ and σ^2 are not known.

Maximum likelihood estimation

- ▶ Maximum likelihood method: Find the parameter value such that the likelihood function is maximized.
- ▶ Let x_1, \dots, x_n be a sample from some population or distribution and let θ be a parameter we want to estimate.
- ▶ The quantity

$$L(x_1, \dots, x_n; \theta) = f(x_1, \dots, x_n; \theta)$$

is called the likelihood function, where f is either the joint PFD or PMF with parameter θ .

- ▶ Since X_1, \dots, X_n are iid, we have

$$\begin{aligned} L(x_1, \dots, x_n; \theta) &= f(x_1, \dots, x_n; \theta) \\ &= f(x_1; \theta) \cdots f(x_n; \theta) = \prod_{i=1}^n f(x_i; \theta) \end{aligned}$$

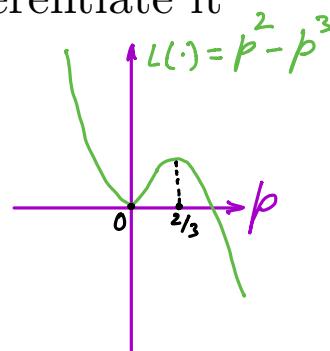
Maximum likelihood estimation

- ▶ Let's try to understand likelihood function and its maximization by example.
- ▶ We want to find the parameter p of the Bernoulli random variable “ $X = 1$ if not defective and $X = 0$ otherwise”.
- ▶ Assume that we have only three samples $x_1 = 1, x_2 = 1, x_3 = 0$. What is the best estimate of p given this information? $\frac{2}{3}$?
- ▶ The likelihood function is

$$L(x_1, x_2, x_3; p) = P(X_1 = 1; p)P(X_2 = 1; p)P(X_3 = 0; p) = p^2(1 - p)$$

- ▶ To find the maximum of the likelihood function, differentiate it and equate to zero:

differentiate:
$$\begin{aligned} \frac{d}{dp} L(x_1, x_2, x_3; p) &= \frac{d}{dp} (p^2 - p^3) \\ &= 2p - 3p^2 \end{aligned}$$



Maximum likelihood estimation

equate to zero: $2\hat{p} - 3\hat{p}^2 = 0$ \hat{p} is the estimate

$$\Rightarrow \hat{p} = 0 \text{ or } \hat{p} = \frac{2}{3}.$$

- To verify that the value indeed maximizes the likelihood function, check whether the second derivative at the estimated value is negative.
 $\Rightarrow \hat{p} = 0 \text{ does not maximize } L(\cdot)$

second derivative $\left. \frac{d^2}{dp^2} L(x_1, x_2, x_3; \theta) \right|_{p=\hat{p}} = 2 - 6\hat{p} \Big|_{p=0} = \underbrace{2}_{>0} > 0 \quad L(\cdot)$

test: $\left. \frac{d^2}{dp^2} L(x_1, x_2, x_3; \theta) \right|_{p=\hat{p}} = 2 - 6\hat{p} \Big|_{p=\frac{2}{3}} = \underbrace{-2}_{<0} < 0, \Rightarrow \hat{p} = \frac{2}{3} \text{ maximizes } L(\cdot).$

- 2/3 is indeed a reasonable estimate of the parameter from the given sample: on average 2 are not defective out of 3.

Maximum likelihood estimation

- ▶ Basic idea of maximum likelihood estimation: the reasonable estimator of a parameter based on a sample is that parameter value that produces the largest probability of obtaining the sample.
- ▶ Definition: Given a random sample (or independent observations) x_1, x_2, \dots, x_n from a PDF/PMF $f(x; \theta)$, the **maximum likelihood estimator** $\hat{\theta}$ is that which maximizes the likelihood function

$$L(x_1, \dots, x_n; \theta) = f(x_1; \theta) \cdots f(x_n; \theta) = \prod_{i=1}^n f(x_i; \theta).$$

Maximum likelihood estimation

generalization of x_1, x_2, x_3 .

- Example: Given a sample x_1, \dots, x_n find maximum likelihood estimator of p for $X \sim \text{Bern}(p)$.

For a Bernoulli r.v. X_i with $P(X_i=1)=p$ and $P(X_i=0)=1-p$, we can write:

$$P(X_i=x_i) = p^{x_i} (1-p)^{1-x_i}, \text{ for } x_i=1, 0.$$

Hence, the likelihood function is:

$$L(x_1, \dots, x_n; p) = \prod_{i=1}^n f(x_i; p)$$

$$= \prod_{i=1}^n P(X_i=x_i; p)$$

$$= \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum x_i} (1-p)^{n-\sum x_i}$$

Maximum likelihood estimation

- Note that the expression is too complex to find \hat{p} .
- If a value maximizes a function then it also maximizes any monotonically increasing function of the function.
- Hence, take log both sides:
$$\underbrace{\log L(x_1, \dots, x_n; p)}_{\text{called Log-Likelihood function}} = \sum x_i \log p + (n - \sum x_i) \log(1-p)$$
- Differentiate:

$$\Rightarrow \frac{d}{dp} \log L(x_1, \dots, x_n; p) = \frac{\sum x_i}{p} - \frac{(n - \sum x_i)}{1-p}$$

Maximum likelihood estimation

- Equate to zero:

$$\Rightarrow \frac{\sum x_i}{\hat{p}} - \frac{(n - \sum x_i)}{1 - \hat{p}} = 0$$

$$\Rightarrow \sum x_i - \sum x_i \hat{p} = n \hat{p} - \sum x_i \hat{p}$$
$$\Rightarrow \hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Second derivative test:

$$\left. \frac{d^2}{dp^2} \ln L(x_1, \dots, x_n; p) \right|_{p=\hat{p}} = \left. \frac{-\sum x_i}{p^2} - \frac{(n - \sum x_i)}{(1-p)^2} \right|_{p=\hat{p}}$$

Maximum likelihood estimation

$$\begin{aligned}
 &= - \frac{\sum x_i}{\left(\frac{\sum x_i}{n}\right)^2} - \frac{(n - \sum x_i)}{\left(\frac{n - \sum x_i}{n}\right)^2} \\
 &= - \frac{n^2}{\sum x_i} - \frac{n^2}{n - \sum x_i} \\
 &= - \left(\frac{n^2}{\sum x_i} + \frac{n^2}{n - \sum x_i} \right)
 \end{aligned}$$

$\Rightarrow \hat{p} = \frac{\sum x_i}{n}$ maximizes the log-likelihood function.

< 0 . $\left(\because \sum_{i=1}^n x_i \text{ can only take values } 0, 1, \dots, n \right)$

$\Rightarrow \hat{p}$ is the maximum likelihood estimate.

Lecture 30: Estimation - Part III

Satyajit Thakor
IIT Mandi

7 August, 2020

Recall: maximum likelihood estimation

- ▶ How to find maximum likelihood estimation of a parameter?
- ▶ Step 1: Write the expression for the likelihood function for the given distribution.
- ▶ Step 2: If the expression is complex, take logarithm both sides.
- ▶ Step 3: Differentiate the likelihood function (or log-likelihood function) with respect to the parameter to be estimated.
- ▶ Step 4: Equate it to zero to find the estimate(s) of the parameter which maximizes the likelihood function.
- ▶ Step 5: Conduct the second derivative test - if the second derivative at an estimate obtained in Step 4 is negative then the estimate provides a (local) maximum value of the function.

Maximum likelihood estimation

- ▶ Example: Given a sample x_1, \dots, x_n find maximum likelihood estimator of λ for $X \sim \text{Pois}(\lambda)$. (Recall Poisson distribution: Lecture 9)

- Recall: $X \sim \text{Pois}(\lambda) : P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!}, x=0,1,2,\dots$

- Then, likelihood function is

$$L(x_1, \dots, x_n; \lambda) = \prod_{i=1}^n f(x_i; \lambda)$$

$$= \prod_{i=1}^n P(X_i = x_i)$$

$$= \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod x_i!}$$

Maximum likelihood estimation

- The log-likelihood function is

$$\begin{aligned}\log_e L(x_1, \dots, x_n; \lambda) &= \log_e \left(\frac{e^{-n\lambda} \times \sum x_i}{\prod x_i!} \right) \\ &= -n\lambda + \sum x_i \log_e \lambda - \log_e (\prod x_i!).\end{aligned}$$

- Differentiate:

$$\frac{d}{d\lambda} \log_e L(x_1, \dots, x_n; \lambda) = -n + \frac{\sum x_i}{\lambda} - 0.$$

- Equate to zero: $-n + \frac{\sum x_i}{\lambda} = 0 \Rightarrow \lambda = \frac{\sum_{i=1}^n x_i}{n}.$

Maximum likelihood estimation

→ Second derivative test :

$$\frac{d^2}{d\lambda^2} \log_e L(x_1, \dots, x_n; \lambda) \Big|_{\lambda=\hat{\lambda}} = - \frac{\sum x_i}{\lambda^2} \Bigg|_{\lambda=\frac{\sum x_i}{n}}$$

$$= - \frac{\sum x_i}{(\sum x_i)^2/n^2}$$

$$= - \frac{n^2}{\sum_{i=1}^n x_i} < 0.$$

$\Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^n x_i}{n}$ maximizes the log-likelihood function
and hence it is the maximum likelihood estimate.

Interval estimation

- ▶ So far we studied point estimation of a parameter.
- ▶ For example, sample mean, sample variance, unbiased point estimators, maximum likelihood estimators.
- ▶ A point estimate \bar{x} of the mean for a given sample x_1, \dots, x_n is not always the parameter μ it estimates but it is “very close” to μ .
- ▶ Hence, rather than a point estimate, it is sometimes more valuable to be able to specify an interval for which we have a certain degree of confidence that μ lies within.
- ▶ To obtain such an interval estimator, we make use of the probability distribution of the point estimator \bar{X} .

Interval estimation

- ▶ An interval estimate of a population parameter θ is an interval of the form $\hat{\theta}_L < \theta < \hat{\theta}_U$, where $\hat{\theta}_L$ and $\hat{\theta}_U$ depend on the value (e.g., $\hat{\theta}$) of the statistic $\hat{\Theta}$ for a particular sample and also on the distribution of the parameter $\hat{\Theta}$.
- ▶ If, for instance, we find $\hat{\theta}_L$ and $\hat{\theta}_U$ such that

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha,$$

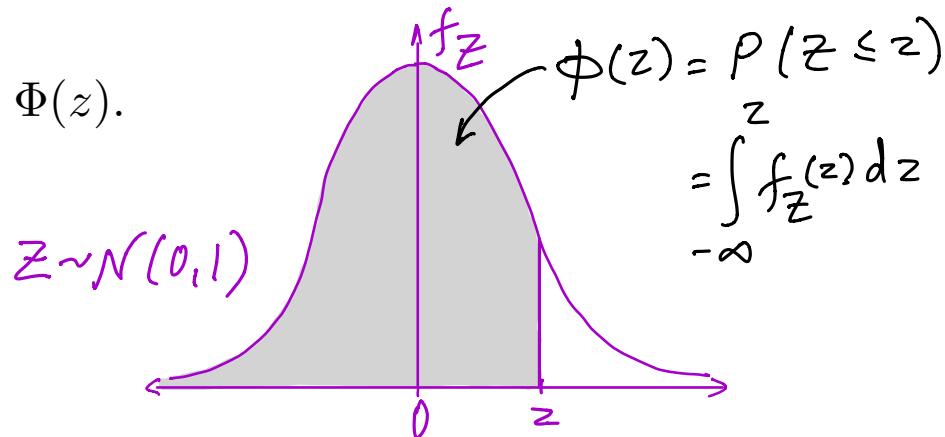
Lower *Upper*

for $0 < \alpha < 1$, then we have a probability of $1 - \alpha$ of selecting a random sample that will produce an interval containing θ .

- ▶ In this case, the interval $\hat{\theta}_L < \theta < \hat{\theta}_U$ is called $100(1 - \alpha)$ percent confidence interval estimate of θ .
- ▶ Next we will try to find $100(1 - \alpha)$ percent confidence interval estimate of a sample mean \bar{X} .

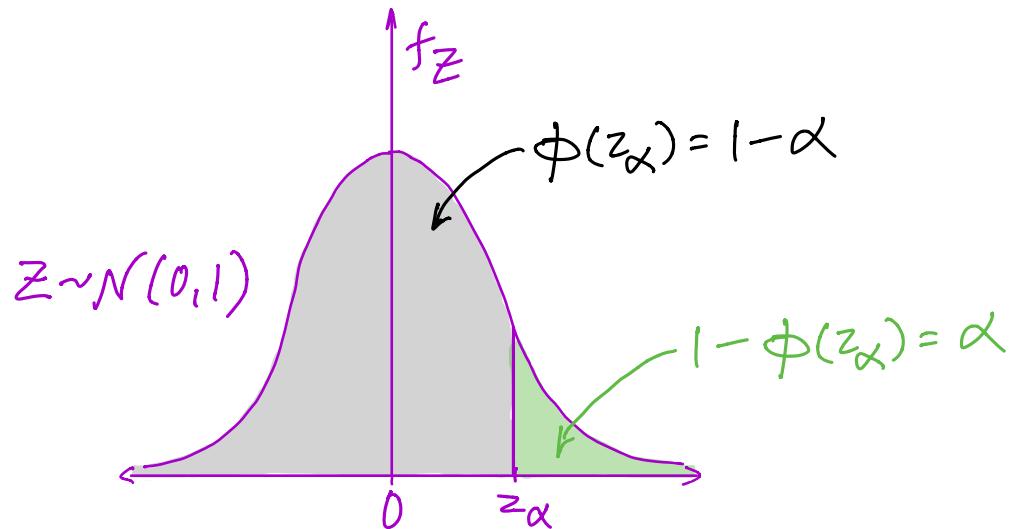
Interval estimation

- ▶ For an example of $100(1 - \alpha)$ percent confidence interval estimate of a sample mean \bar{X} , assume that \bar{X} is normally distributed.
- ▶ Note that this is a very logical assumption for large n since the CLT suggests that the distribution of \bar{X} can be well approximated by the normal distribution $\mathcal{N}(\mu, \sigma^2/n)$ (recall: Slides 4-5 of Lecture 26).
- ▶ Also, recall that if \bar{X} has the distribution $\mathcal{N}(\mu, \sigma^2/n)$ then $\bar{X} = \mu + (\sigma/\sqrt{n})Z$, where, $Z \sim \mathcal{N}(0, 1)$ (recall: Slide 7 of Lecture 25).
- ▶ Now, denote $P(Z \leq z) = \Phi(z)$.



Interval estimation

- ▶ Let z_α be the value such that $\Phi(z_\alpha) = 1 - \alpha$.
- ▶ For any give $0 < \alpha < 1$, we can find out z_α numerically (using the CDF table for standard normal).



Lecture 31: Estimation - Part IV

Satyajit Thakor
IIT Mandi

11 August, 2020

Recall - Interval estimation

- ▶ An interval estimate of a population parameter θ is an interval of the form $\hat{\theta}_L < \theta < \hat{\theta}_U$, where $\hat{\theta}_L$ and $\hat{\theta}_U$ depend on the value (e.g., $\hat{\theta}$) of the statistic $\hat{\Theta}$ for a particular sample and also on the distribution of the parameter $\hat{\Theta}$.
- ▶ If, for instance, we find $\hat{\theta}_L$ and $\hat{\theta}_U$ such that

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha,$$

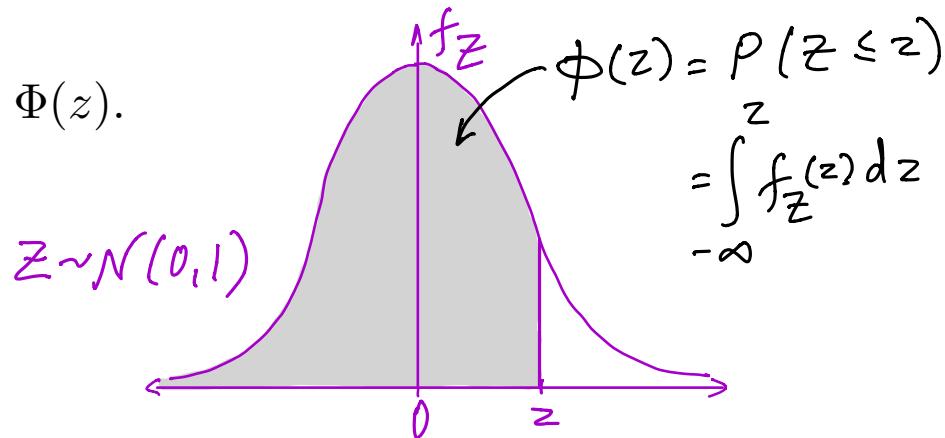
Lower *Upper*

for $0 < \alpha < 1$, then we have a probability of $1 - \alpha$ of selecting a random sample that will produce an interval containing θ .

- ▶ In this case, the interval $\hat{\theta}_L < \theta < \hat{\theta}_U$ is called $100(1 - \alpha)$ percent confidence interval estimate of θ .
- ▶ Next we will try to find $100(1 - \alpha)$ percent confidence interval estimate of a sample mean \bar{X} .

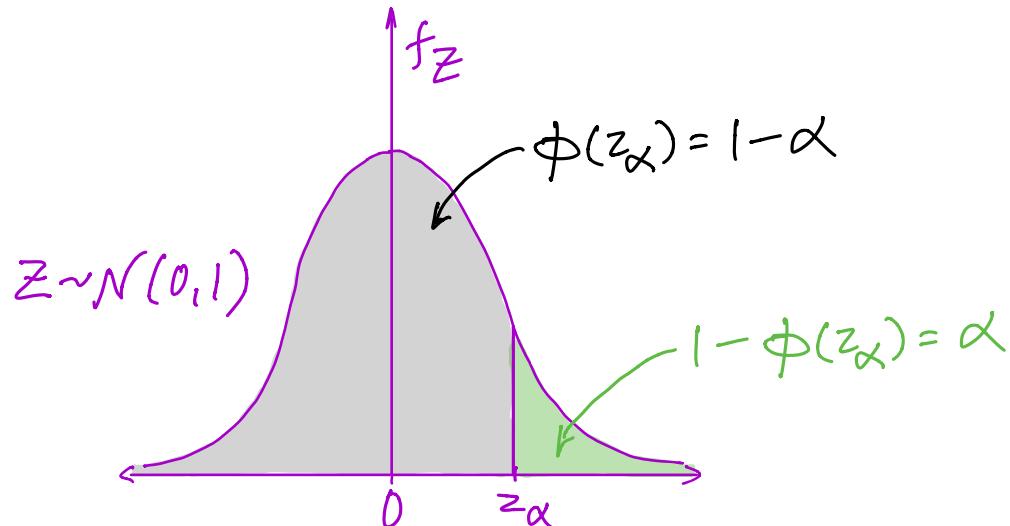
Recall - Interval estimation

- ▶ For an example of $100(1 - \alpha)$ percent confidence interval estimate of a sample mean \bar{X} , assume that \bar{X} is normally distributed.
- ▶ Note that this is a very logical assumption for large n since the CLT suggests that the distribution of \bar{X} can be well approximated by the normal distribution $\mathcal{N}(\mu, \sigma^2/n)$ (recall: Slides 4-5 of Lecture 26).
- ▶ Also, recall that if \bar{X} has the distribution $\mathcal{N}(\mu, \sigma^2/n)$ then $\bar{X} = \mu + (\sigma/\sqrt{n})Z$, where, $Z \sim \mathcal{N}(0, 1)$ (recall: Slide 7 of Lecture 25).
- ▶ Now, denote $P(Z \leq z) = \Phi(z)$.



Recall - Interval estimation

- ▶ Let z_α be the value such that $\Phi(z_\alpha) = 1 - \alpha$.
- ▶ For any give $0 < \alpha < 1$, we can find out z_α numerically (using the CDF table for standard normal).

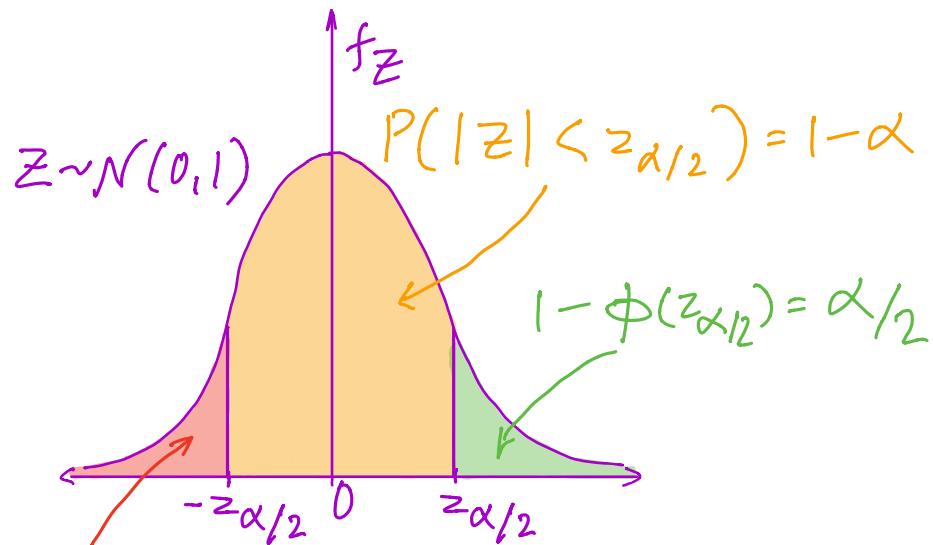


Interval estimation

► $P(|Z| < z_{\alpha/2}) = 1 - \alpha$

► Proof:

$$\begin{aligned} P(|Z| < z_{\alpha/2}) &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\ &= \phi(z_{\alpha/2}) - \phi(-z_{\alpha/2}) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} \\ &= 1 - \alpha. \end{aligned}$$



$$\underbrace{\phi(-z_{\alpha/2})}_{=} = \alpha/2$$

$\because f_Z$ is symmetric around 0.

$$\Rightarrow \phi(-z_{\alpha/2}) = 1 - \phi(z_{\alpha/2}) = \alpha/2.$$

Interval estimation

- For our sample mean example, we have $\bar{X} = \mu + (\sigma/\sqrt{n})Z$ where Z is a standard normal.

$$\text{Then, } P(|Z| < z_{\alpha/2}) = 1 - \alpha$$

$$\Rightarrow P(|Z| < z_{\alpha/2})$$

$$= P(-z_{\alpha/2} < Z < z_{\alpha/2})$$

$$= P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right)$$

$$= P\left(-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Interval estimation

- ▶ Thus we have established the following result for interval estimate of a sample mean:
- ▶ If \bar{x} is the mean of a random sample of size n from a population with known variance σ^2 , a $100(1 - \alpha)$ percent confidence interval for μ is given by

$$\hat{\theta}_L < \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \hat{\theta}_U$$

where $z_{\alpha/2}$ is the value such that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$

Interval estimation

- ▶ Example: The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per milliliter. Find the 95% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 gram per milliliter.

- Let X be the r.v.: "zinc concentration."
- Given: sample mean $\bar{x} = 2.6 \text{ grams/ml}$ & $\sigma_x = 0.3 \text{ grams/ml}$.
- \bar{x} is an estimate of μ_X .
- For 95% confidence interval $\underbrace{\alpha = 0.05}_{\therefore 95 = 100(1-\alpha)}$

Interval estimation

- Then, recall that, $P(Z < z_{\alpha/2}) = 1 - \alpha/2$
 $\Rightarrow \Phi(z_{0.025}) = 1 - 0.025$
 $= 0.975$
- By the CDF table, $z_{0.025} = 1.96$.
- The $100(1-\alpha)\%$. confidence interval of μ is
$$\bar{x} - z_{\alpha/2} \cdot \frac{6}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{6}{\sqrt{n}}$$
- Hence, 95% confidence interval of μ is
$$2.6 - (1.96) \cdot \frac{0.3}{\sqrt{36}} < \mu_x < 2.6 + (1.96) \cdot \frac{0.3}{\sqrt{36}}$$

$$\Rightarrow 2.50 < \mu_x < 2.69$$

Lecture 32:
Monte Carlo Method
&
Hypothesis Testing - Part I

Satyajit Thakor
IIT Mandi

14 August, 2020

Monte Carlo method

- ▶ Monte Carlo method is a set of algorithms which use randomness to solve problems either exactly or approximately.
- ▶ We will study Monte Carlo integration - a method used to approximate a definite integral using random point generation.
- ▶ Consider the problem of approximating the definite integral

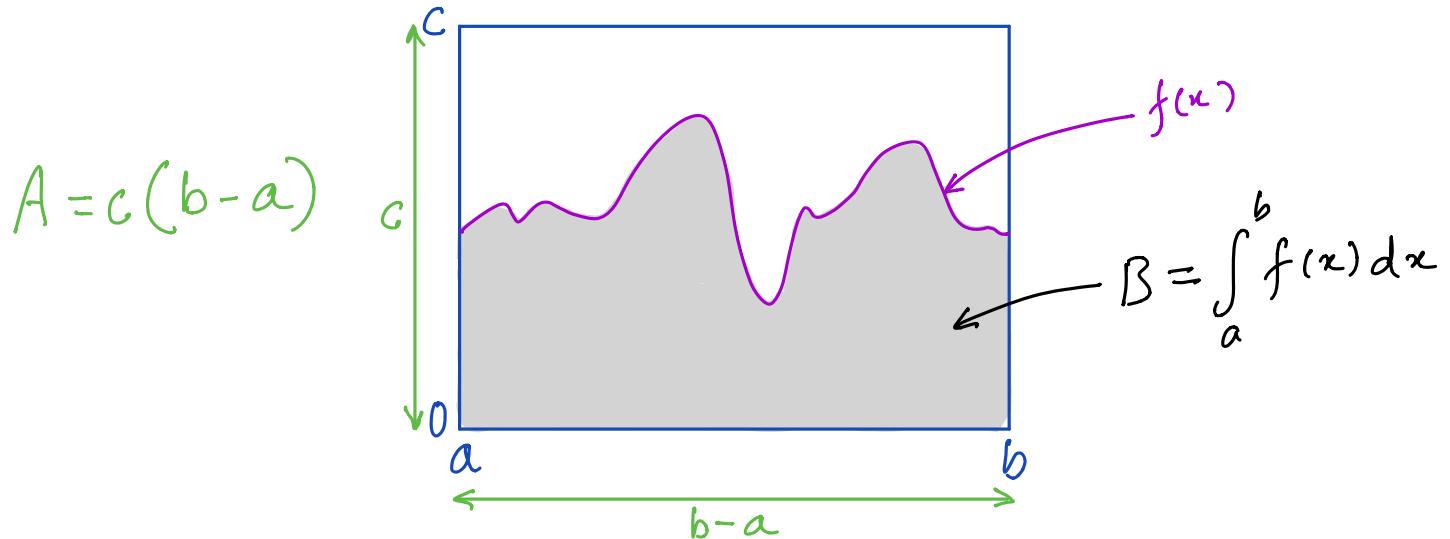
$$\int_a^b f(x)dx$$

where, f is some complicated function so that exact integration is not possible with any existing method.

- ▶ Assume that $0 \leq f(x) \leq c$ and hence the integral is finite.
- ▶ We can generate and use random numbers to approximate this integral!

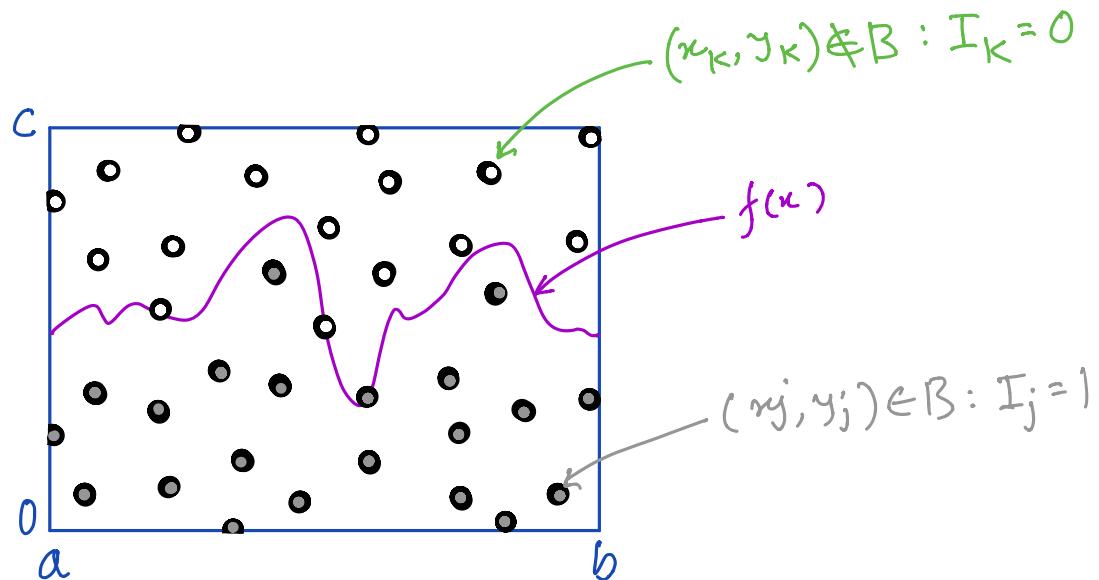
Monte Carlo method

- ▶ We can generate and use random numbers to approximate this integral!
- ▶ Let A be the area of the rectangle in the (x, y) -plane defined by $a \leq x \leq b$ and $0 \leq y \leq c$.
- ▶ Let B be the region under the curve $y = f(x)$ for $a \leq x \leq b$,
- ▶ Thus, the area of B is the desired integral.



Monte Carlo method

- Method: take random samples from A , then calculate the proportion of the samples that also fall into the area B .
- Generate iid points $(X_1, Y_1), \dots, (X_n, Y_n)$ uniformly over A .
Now let I_j be the Bernoulli r.v. such that $I_j = 1$ if $(X_j, Y_j) \in B$ and 0 otherwise.



Monte Carlo method

- Then

$$p = E(I_j) = P(I_j = 1) = \frac{B}{A} = \frac{\int_a^b f(x)dx}{c(b-a)}.$$

- By the WLLN, for large n we can approximate p as

$$\frac{1}{n} \sum_{j=1}^n I_j.$$

- Hence,

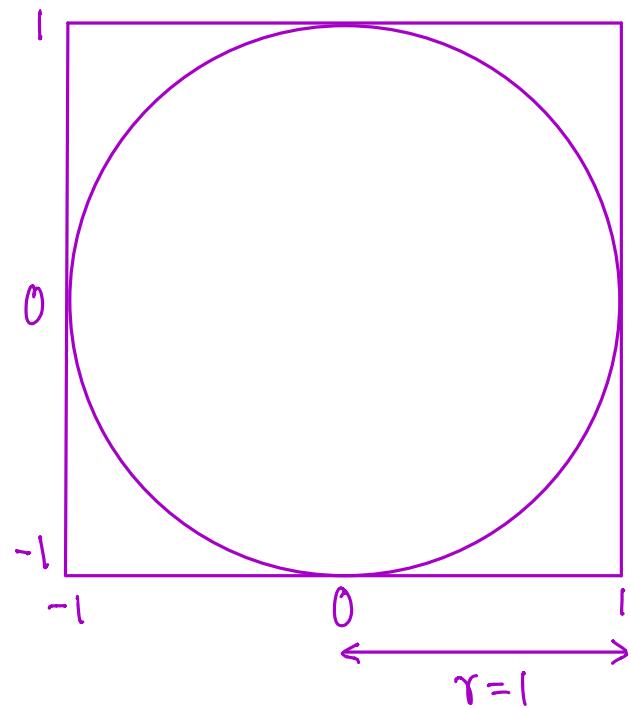
$$\frac{1}{n} \sum_{j=1}^n I_j \approx \frac{\int_a^b f(x)dx}{c(b-a)}$$

$$\Rightarrow \int_a^b f(x)dx \approx \frac{c(b-a)}{n} \sum_{j=1}^n I_j$$

Monte Carlo method

- ▶ Example: Approximate the value of the number π .

- Generate samples $(x_1, y_1), \dots, (x_n, y_n)$ iid uniformly in the square.



- Let $I_j = \begin{cases} 1 & \text{if } x_j^2 + y_j^2 \leq 1 \\ 0 & \text{otherwise.} \end{cases}$

- Then, $\pi r^2 = \pi \approx 4 \cdot \frac{1}{n} \sum_{j=1}^n I_j$
area of the square

Hypothesis testing

- ▶ We, the scientists/engineers do not only come across problems of point/interval estimation in practice.
- ▶ Often, given a sample, we need to decide whether certain “statement” is true.
- ▶ For example, a medical researcher may decide on the basis of experimental evidence whether coffee drinking increases the risk of cancer in humans.
- ▶ Here “coffee drinking increases the risk of cancer in humans” is a statement or conjecture.
- ▶ A conjecture is a statement which a scientist proposes but correctness of which is yet to be established.
- ▶ A data scientist attempts to “test” correctness of the statement based on the available sample/evidence.

Hypothesis testing

- ▶ A **statistical hypothesis** is a conjecture concerning a population.
- ▶ Whether a hypothesis is true or false is never known with absolute certainty unless we examine the entire population or underlying distribution.
- ▶ We take a random sample from the population of interest and use the data contained in this sample to provide evidence that either supports or does not support the hypothesis.
- ▶ Evidence from the sample that is inconsistent with the stated hypothesis leads to a rejection or nullification of the hypothesis.

Hypothesis testing

- ▶ Another simple story to understand hypothesis testing: A person is arrested based on a suspicion of committing a crime.
- ▶ The court has the following hypothesis to test: The person is innocent.
- ▶ The goal of the court is to nullify this hypothesis based on evidence.
- ▶ In the court, evidence is presented and examined.
- ▶ Based on the evidence, the jury either fails to reject the hypothesis (the person is innocent) or rejects it (the person is guilty of the crime).
- ▶ Important thing to note:
- ▶ If the jury fails to reject the hypothesis based on the evident presented, it does imply that the person is innocent.
- ▶ It only implies that the evidence was insufficient to convict.

Lecture 33: Hypothesis Testing - Part II

Satyajit Thakor
IIT Mandi

17 August, 2020

Hypothesis testing

- ▶ Any hypothesis we wish to test is called **null hypothesis** and is denoted by H_0 .
- ▶ The rejection of H_0 leads to the acceptance of an **alternative hypothesis**, denoted by H_1 .
- ▶ The null hypothesis H_0 nullifies or opposes H_1 and is often the logical complement to H_1 .
- ▶ After analysis of a sample, the analyst arrives at one of the following two conclusions:
- ▶ (1) reject H_0 in favor of H_1 because of sufficient evidence in the data or
- ▶ (2) fail to reject H_0 because of insufficient evidence in the data.

Hypothesis testing

- ▶ How to test a hypothesis based on a sample?
- ▶ Let's understand this by example: A certain type of flu vaccine is known to be only 25% effective after a period of 2 years. To determine if a new and somewhat more expensive vaccine is superior in providing protection against the same virus for a longer period of time, suppose that 20 people are chosen at random and administered the vaccine dose.
- ▶ If more than 8 of those receiving the new vaccine surpass the 2-year period without contracting the virus, the new vaccine will be considered superior to the one presently in use.
- ▶ Let X be the binomial r.v. $\text{Bin}(20, p)$: “number of people who receive protection for a period of at least 2 years”

Hypothesis testing

- ▶ We are essentially testing the null hypothesis that the new vaccine is equally effective after a period of 2 years as the one now commonly used.
- ▶ The alternative hypothesis is that the new vaccine is in fact superior.
- ▶ This is equivalent to testing the hypothesis that the binomial parameter for the probability of a success on a given trial is $p = 1/4$ against the alternative that $p > 1/4$.
- ▶ Then the null and alternative hypothesis are

$$H_0 : p = 0.25,$$

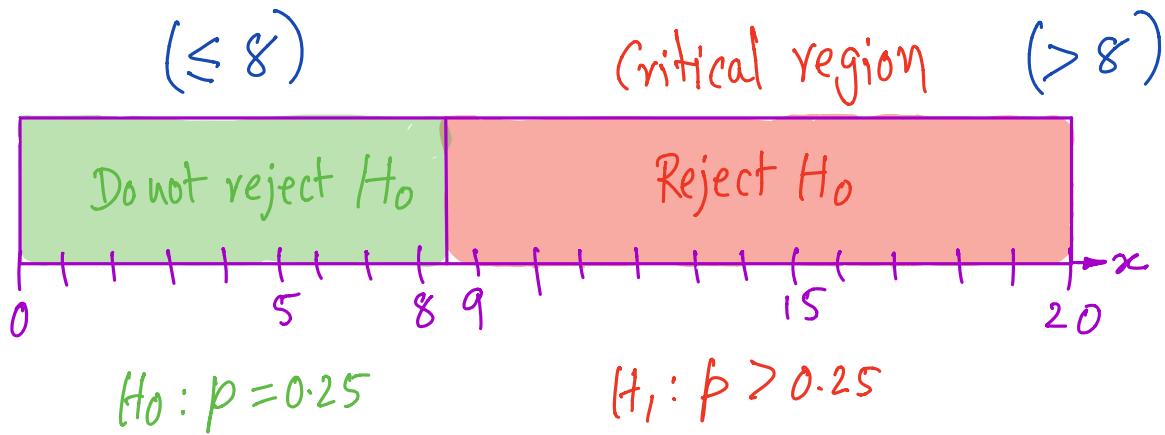
$$H_1 : p > 0.25.$$

- ▶ The possible values of X are from 0 to 20.

Hypothesis testing

defined by the data analyst, it may be defined as 7 or 9 too.

- ▶ If the score is greater than 8 then we reject the null hypothesis H_0 . This is referred to as the critical region or rejection region.
 - ▶ The last value before passing into the critical region is called the critical value
 - ▶ If the score is less than or equal to 8 then we fail to reject the null hypothesis H_0 .



Hypothesis testing

- ▶ Null hypothesis is usually stated as an equality.
- ▶ A test of any statistical hypothesis where the alternative is one sided, such as

$$H_0 : \theta = \theta_0,$$

$$H_1 : \theta > (\text{or } <) \theta_0$$

is called a **one-tailed test**.

- ▶ A test of any statistical hypothesis where the alternative is two sided, such as

$$H_0 : \theta = \theta_0,$$

$$H_1 : \theta \neq \theta_0$$

is called a **two-tailed test**.

Hypothesis testing

- ▶ Example: A manufacturer of a certain brand of snack claims that the average saturated fat content does not exceed 1.5 grams per serving. State the null and alternative hypotheses to be used in testing this claim and determine the critical region.

- Let X be "the saturated fat content in a serving."

$$\begin{array}{l} H_0 : \mu_X = 1.5 \\ H_1 : \mu_X > 1.5 \end{array} \quad \left. \begin{array}{l} \\ \end{array} \right\} \text{one-tailed test}$$

- Assume that the null hypothesis is rejected if $\bar{X} \geq 1.7$.

- Then, this is the one tailed test and the critical region is the interval from 1.7 to $+\infty$. That is, if $\underbrace{\bar{X}}_{\text{Sample mean}} \in [1.7, \infty)$ then we reject H_0 .

Hypothesis testing

- ▶ Example: A real estate agent claims that 60% of all residences being built today are 3-bedroom homes. To test this claim, a large sample of new residences is inspected; the proportion of these homes with 3 bedrooms is recorded and used as the test statistic. State the null and alternative hypotheses to be used in this test and determine the critical region.

- Let $X = 1$ if a home built today is 3-bedroom and 0 otherwise. Then,

$$\begin{aligned} H_0 : \mu_X &= 0.6 \\ H_1 : \mu_X &\neq 0.6 \end{aligned} \quad \left. \begin{array}{l} \\ \end{array} \right\} \text{two-tailed test}$$

- Assume that the sample size is 100. Then the critical region can be defined as $[0, 55]$ and $[65, 100]$ where we reject H_0 if $X_1 + \dots + X_{100}$ is in the critical region.

Hypothesis testing

- ▶ Now, note that we reject or fail to reject H_0 only based on the sample and hence, there may be an error in testing a hypothesis.
- ▶ Rejection of the null hypothesis H_0 when it is true is called a **type I error**.
- ▶ Nonrejection of H_0 when it is false is called a **type II error**.

	H_0 is true	H_0 is false
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Hypothesis testing

- ▶ The probability of committing a type I error, also called the level of significance, is denoted by α .
- ▶ Type I error: Rejection of the null hypothesis when it is true.
- ▶ For our example, the type I error occurs if $X > 8$ but $p = 1/4$.
- ▶ Let's find the probability of the type I error:

$$\begin{aligned}\text{Level of significance } \alpha &= P\{\text{type I error}\} \\ &= P(X > 8 \text{ when } p = 1/4) \\ &= \sum_{k=9}^{20} \binom{20}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{20-k} \\ &= 0.04093\end{aligned}$$

Lecture 34: Hypothesis Testing - Part III

Satyajit Thakor
IIT Mandi

21 August, 2020

Hypothesis testing

- ▶ The probability of committing a type II error, denoted by β , is impossible to compute unless we have a specific alternative hypothesis.
- ▶ Type II error: Nonrejection of the H_0 when it is false.
- ▶ For our example, the type II error occurs if $p > 1/4$ when $X \leq 8$.
- ▶ Let the particular alternative hypothesis be $p = .5 > 1/4$. Then

$$\begin{aligned}\beta &= P\{\text{type II error}\} \\ &= P(X \leq 8 \text{ when } p = 1/2) \\ &= \sum_{k=0}^{20} \binom{20}{k} (1/2)^{20} = 0.25172\end{aligned}$$

Hypothesis testing

- That is, it is quite likely (type II error prob. is 0.2517) that we shall reject the new vaccine when, in fact, it is superior (50% effective) to what is now in use (compared to 25% effective).
- Now, let the particular alternative hypothesis be $p = .7 > 1/4$. Then,

$$\begin{aligned}\beta &= P\{\text{type II error}\} \\ &= P(X \leq 8 \text{ when } p = .7) \\ &= \sum_{k=0}^{8} \binom{20}{k} (.7)^k (.3)^{20-k} = 0.00514.\end{aligned}$$

- That is, it is extremely unlikely that the new vaccine would be rejected when it was 70% effective after a period of 2 years.

Hypothesis testing

- ▶ How to decrease Type I or Type II error?
- ▶ Let's assume that we want to reduce the possibility of Type II error. This can be done by increasing the critical region:

- Let the critical value be 7.

- Now, we test $p = 1/4$ against $p = 1/2$.

- Then,

$$\alpha = \sum_{k=8}^{20} \binom{20}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{20-k} = 0.1018$$

$$\beta = \sum_{k=0}^7 \binom{20}{k} \left(\frac{1}{2}\right)^{20} = 0.1316$$

- Thus, by increasing the critical region β is reduced from 0.2517 to 0.1316 but α is increased from 0.0409 to 0.1018.

Hypothesis testing

- ▶ How to decrease the probability of both type I and II errors?
- ▶ The probability of committing both types of error can be reduced by increasing the sample size.
 - Consider a sample of size 100.
 - If more than 36 are protected from the virus for 2 years then we reject $H_0 : p = \frac{1}{4}$.
 - Critical region is $37, \dots, 100$.
Recall: CLT (Lec. 26)
 - To compute the errors we use normal approximation.

$$X \sim \text{Bin}(n, p) \approx N\left(\mu_X, \sigma_X^2\right) \sim Y$$

n μ_X n σ_X^2

$$\underbrace{n}_{H_0} \quad \underbrace{100p}_{\mu_Y} \quad \underbrace{100p(1-p)}_{\sigma_Y^2}$$

$$- p = \frac{1}{4} \Rightarrow Y_0 \sim N\left(\underbrace{25}_{\mu_{Y_0}}, \underbrace{18.75}_{\sigma_{Y_0}^2}\right) \quad Y_0 = \mu_{Y_0} + \sigma_{Y_0} Z \sim N(0, 1)$$

Hypothesis testing

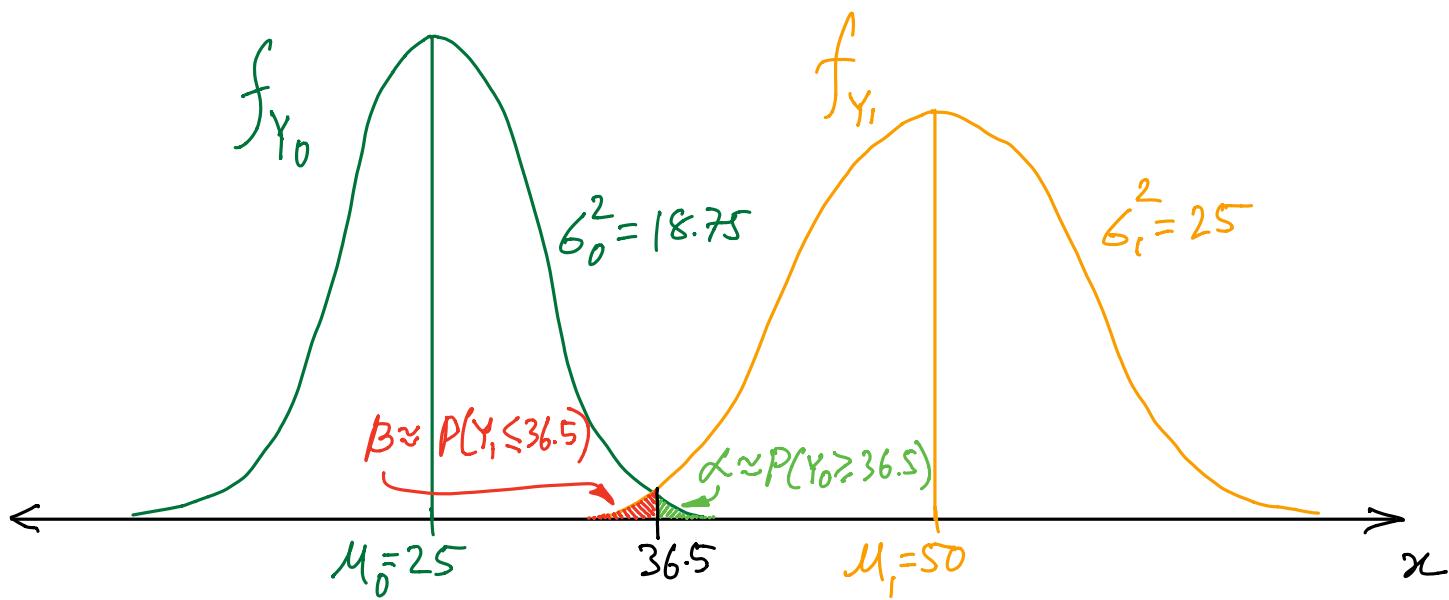
- Then, $\alpha = P\{\text{Type I error}\}$ $= P(X > 36 \text{ when } p=1/4)$ $\approx P(Y_0 \geq 36.5)$ $= P\left(Z \geq \frac{36.5 - 25}{\sqrt{18.75}}\right)$ $= P(Z \geq 2.66)$ $= 1 - \phi_z(2.66) = 1 - 0.9961 = 0.0039.$
- Now, let the alternative hypothesis be $p = 1/2$.
- $\underbrace{p = 1/2}_{H_1} \Rightarrow Y_i \sim N(\underbrace{50}_{\mu_{Y_i}}, \underbrace{25}_{\sigma_{Y_i}^2})$, $Y_i = \mu_{Y_i} + \sigma_{Y_i} Z \sim N(0, 1)$.
- Then, $\beta = P\{\text{Type II error}\}$

Hypothesis testing

$$= P(X \leq 36 \text{ when } p = \mu_2)$$

$$\approx P(Y_1 \leq 36.5)$$

$$= P\left(Z \leq \frac{36.5 - 50}{\sqrt{25}}\right) = P(Z \leq -2.7) = 0.0035$$



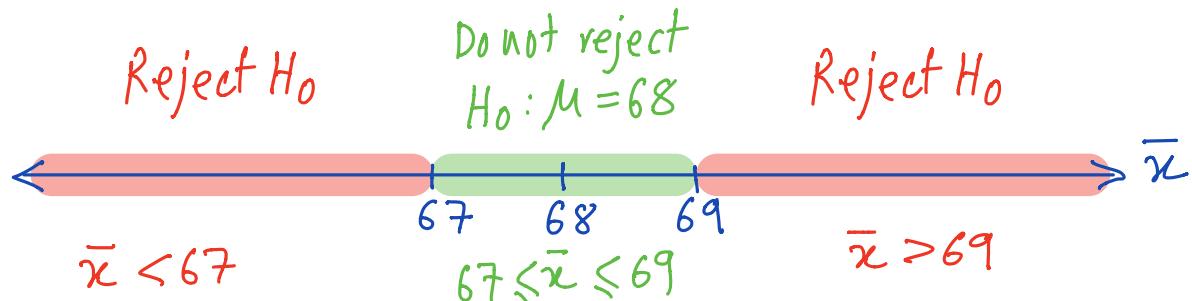
Hypothesis testing

- ▶ Example: Consider the null hypothesis that the average weight of students in a college is 68 kgs against the alternative hypothesis that it is unequal to 68:

$$\begin{aligned} H_0 : \mu &= 68, \\ H_1 : \mu &\neq 68. \end{aligned} \quad \left. \begin{array}{l} \\ \end{array} \right\} \text{two-tailed test}$$

- ▶ Assume that the weight is a normally distributed with $\sigma = 3.6$ and the sample size is 36.
 - (a) Define a suitable critical region
 - (b) Find the probability of Type I error
 - (c) Find the probability of Type II error
- (a) We define the critical region as: $\bar{x} < 67$, $\bar{x} > 69$.
- That is, do not reject if $67 \leq \bar{x} \leq 69$ and reject otherwise.

Hypothesis testing



(b) Note that, $n = 36$ (sample size)

- Let X be "Weight of a student"

- Let \bar{X} be the sample mean. Then,

$$\mu_{\bar{X}} = \mu_X, \quad \sigma_{\bar{X}}^2 = \sigma_X^2 / n = 3.6^2 / 36 = 0.36.$$

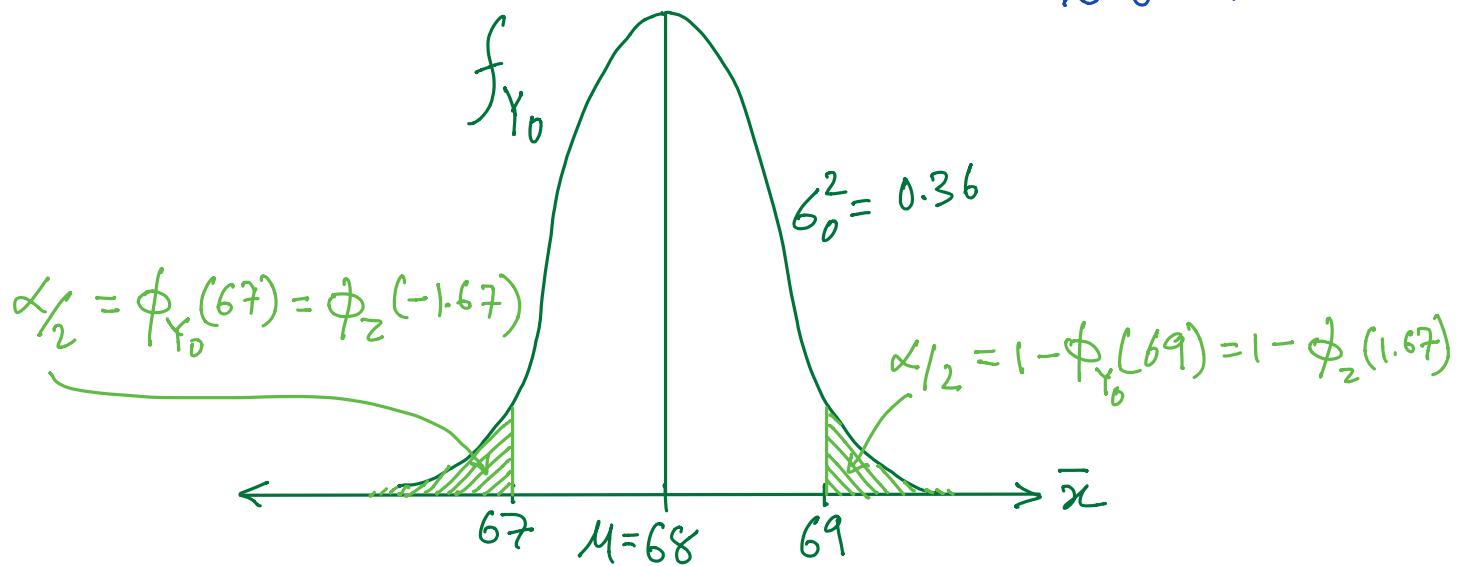
- Then, $\alpha = P(\bar{X} < 67 \text{ when } \mu = 68)$

$$+ P(\bar{X} > 69 \text{ when } \mu = 68)$$

Hypothesis testing

Let: $Y_0 \sim N(68, 0.36)$ $Y_0 = 68 + 0.6Z$

$$\begin{aligned}\Rightarrow \alpha &= P(Y_0 < 67) + P(Y_0 > 69) \\ &= P(Z < -1.67) + P(Z > 1.67) = 2 \phi_z(-1.67) \\ &\approx 0.0950.\end{aligned}$$



Lecture 35: Hypothesis Testing - Part IV

Satyajit Thakor
IIT Mandi

24 August, 2020

Recall

- ▶ Example: Consider the null hypothesis that the average weight of students in a college is 68 kgs against the alternative hypothesis that it is unequal to 68:

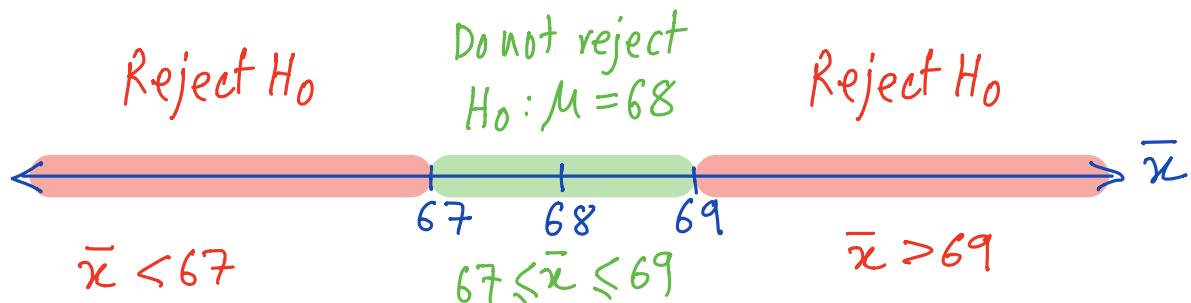
$$\begin{aligned} H_0 : \mu &= 68, \\ H_1 : \mu &\neq 68. \end{aligned} \quad \left. \begin{array}{l} \\ \end{array} \right\} \text{two-tailed test}$$

- ▶ Assume that the weight is a normally distributed with $\sigma = 3.6$ and the sample size is 36.
 - (a) Define a suitable critical region
 - (b) Find the probability of Type I error
 - (c) Find the probability of Type II error

(a) We define the critical region as: $\bar{x} < 67, \bar{x} > 69$.

- That is, do not reject if $67 \leq \bar{x} \leq 69$ and reject otherwise.

Recall



(b) Note that, $n = 36$ (sample size)

- Let X be "Weight of a student"

- Let \bar{X} be the sample mean. Then,

$$\mu_{\bar{X}} = \mu_X, \quad \sigma_{\bar{X}}^2 = \sigma_X^2/n = 3.6^2/36 = 0.36.$$

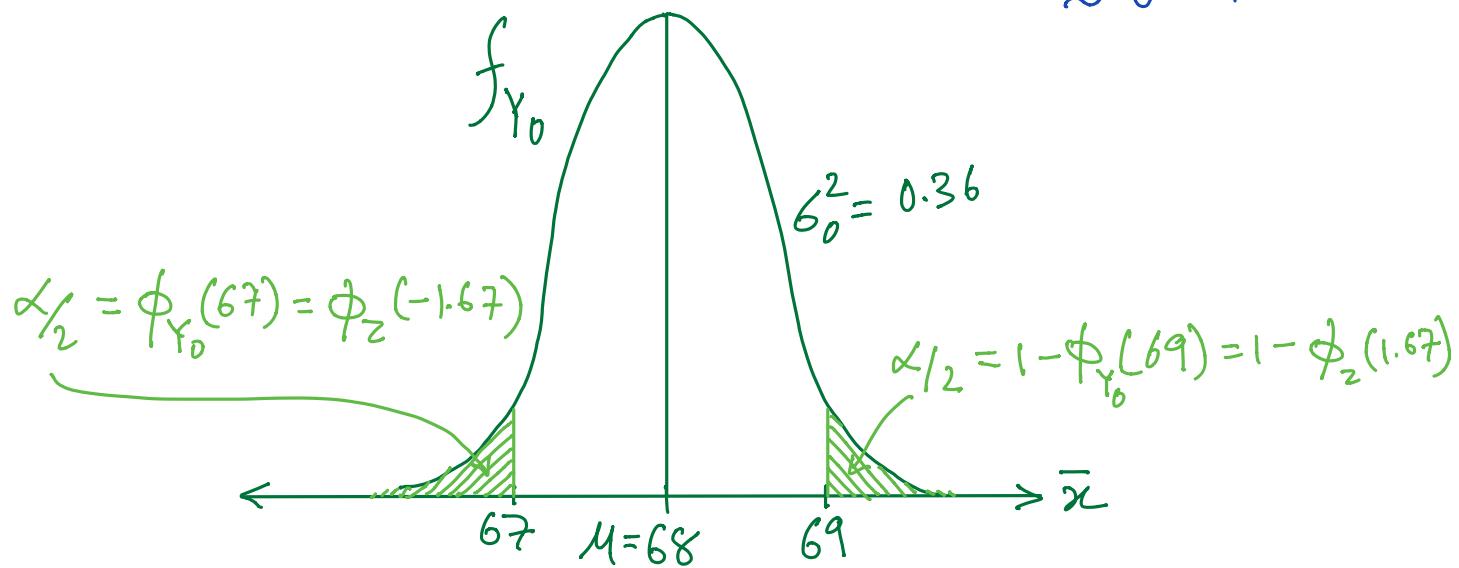
- Then, $\alpha = P(\bar{X} < 67 \text{ when } \mu = 68)$

$$+ P(\bar{X} > 69 \text{ when } \mu = 68)$$

Recall

$$\rightarrow \text{Let: } Y_0 \sim N(68, 0.36) \quad Y_0 = 68 + 0.6Z$$

$$\begin{aligned}\Rightarrow \alpha &= P(Y_0 < 67) + P(Y_0 > 69) \\ &= P(Z < -1.67) + P(Z > 1.67) = 2 \Phi_Z(-1.67) \\ &\approx 0.0950.\end{aligned}$$



Hypothesis testing

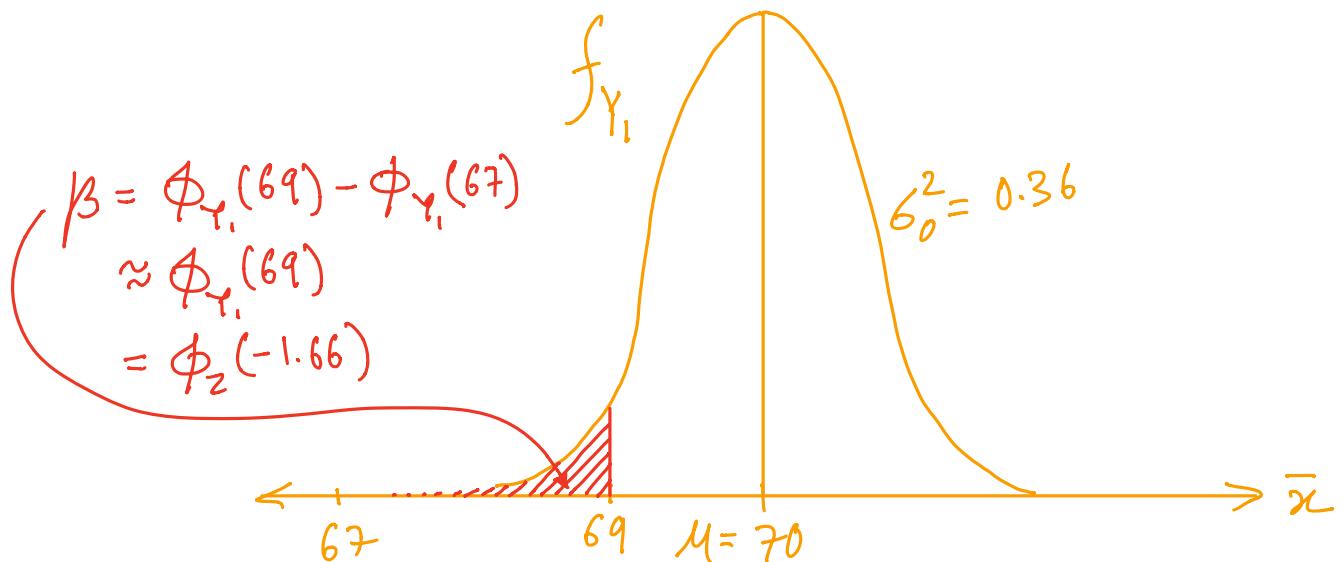
(c) To compute β , let the alternative hypothesis be $\mu = 70$.

- Then, $\beta = P(67 \leq \bar{X} \leq 69 \text{ when } \mu = 70)$

- Let: $Y_i \sim N(70, 0.36)$, $Y_i = 70 + 0.6Z$.

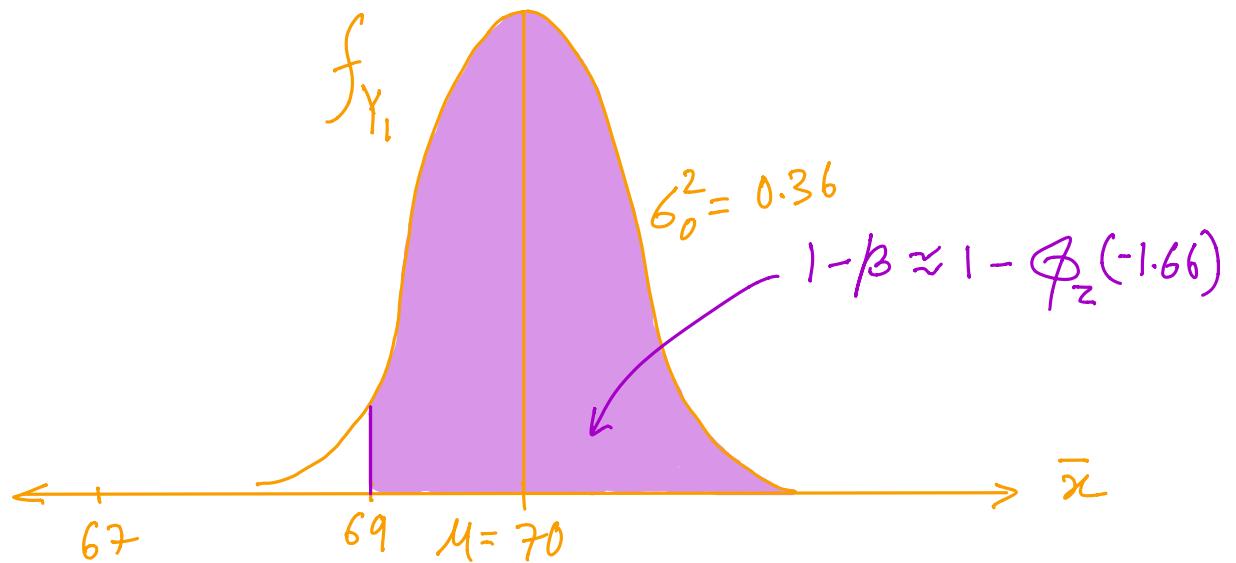
$$\begin{aligned}\Rightarrow \beta &= P(67 \leq Y_i \leq 69) \\ &= P\left(-\frac{3}{.6} \leq Z \leq -\frac{1}{.6}\right) \\ &= \Phi_Z(-1.66) - \Phi_Z(-5) \\ &\approx 0.0485 - 0\end{aligned}$$

Hypothesis testing



Hypothesis testing

- The power of a test is the probability of rejecting H_0 given that a specific alternative is true, i.e., $1 - \beta$.



Hypothesis testing

- ▶ In general, one wants to control the probability of committing type I or type II error or both.
 - ▶ It is customary to choose $\alpha = .05$, or in some tests, $\alpha = .01$.
 - ▶ p -value is an important parameter to decide whether to reject the null hypothesis.
- ▶ **p -values** are the probability of obtaining an effect at least as extreme as the one in your sample data, assuming the truth of the null hypothesis.
- ▶ Question: If the p -value for a given sample is very low, shall we accept or reject H_0 ?

- Answer: reject ! (why?: let's understand by example)

Hypothesis testing

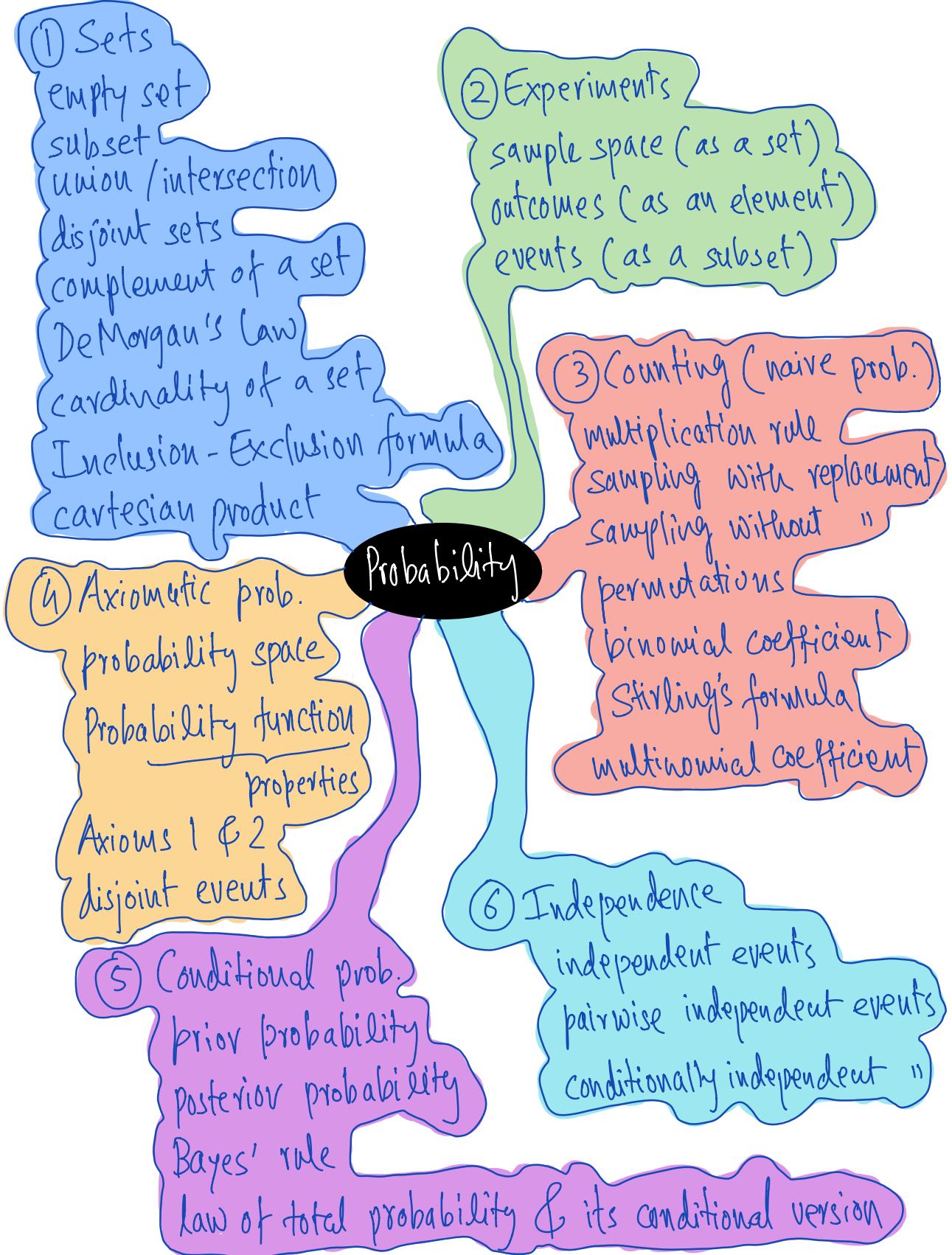
- Example: Suppose the observed value of \bar{X} is 69.3.

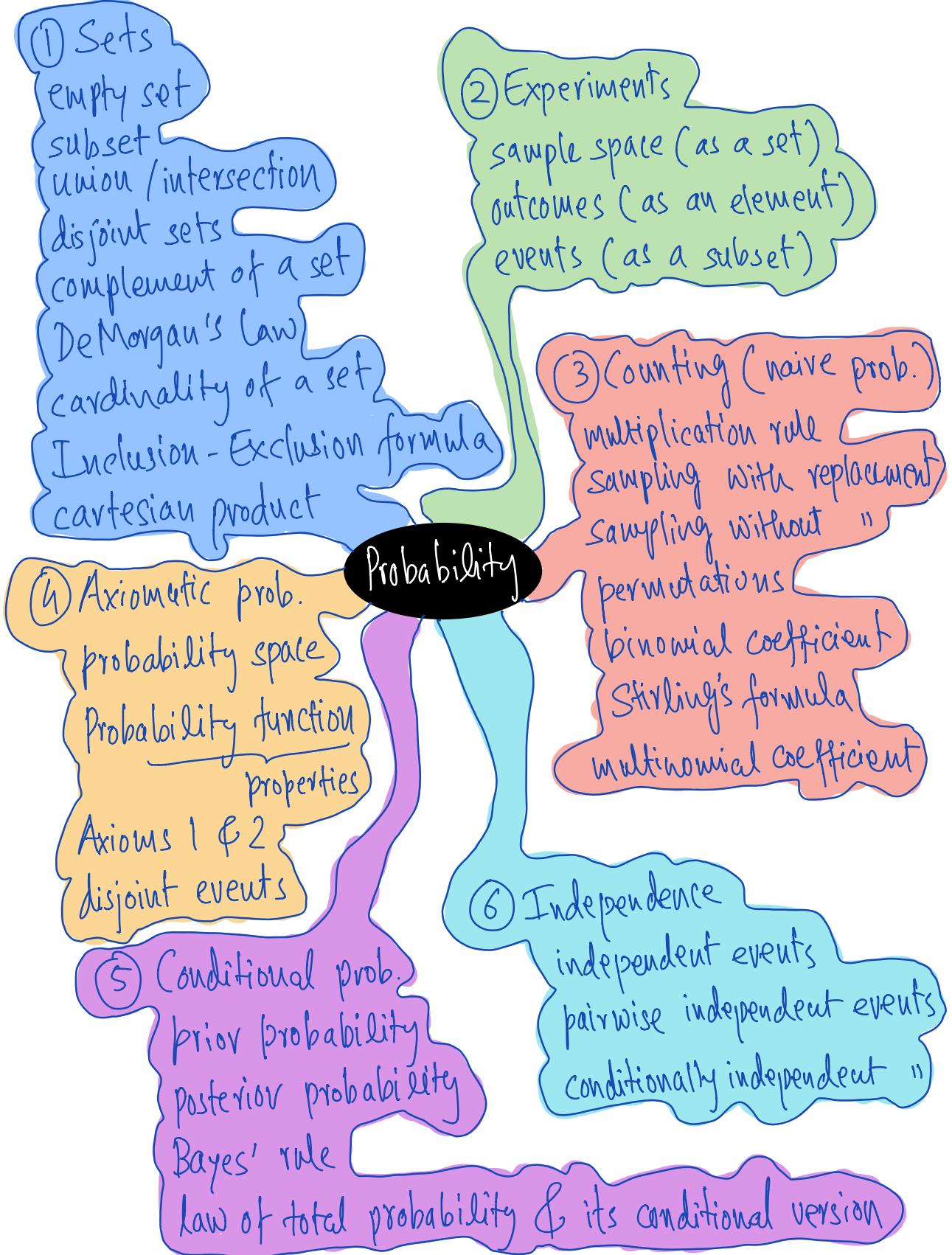
- Then,

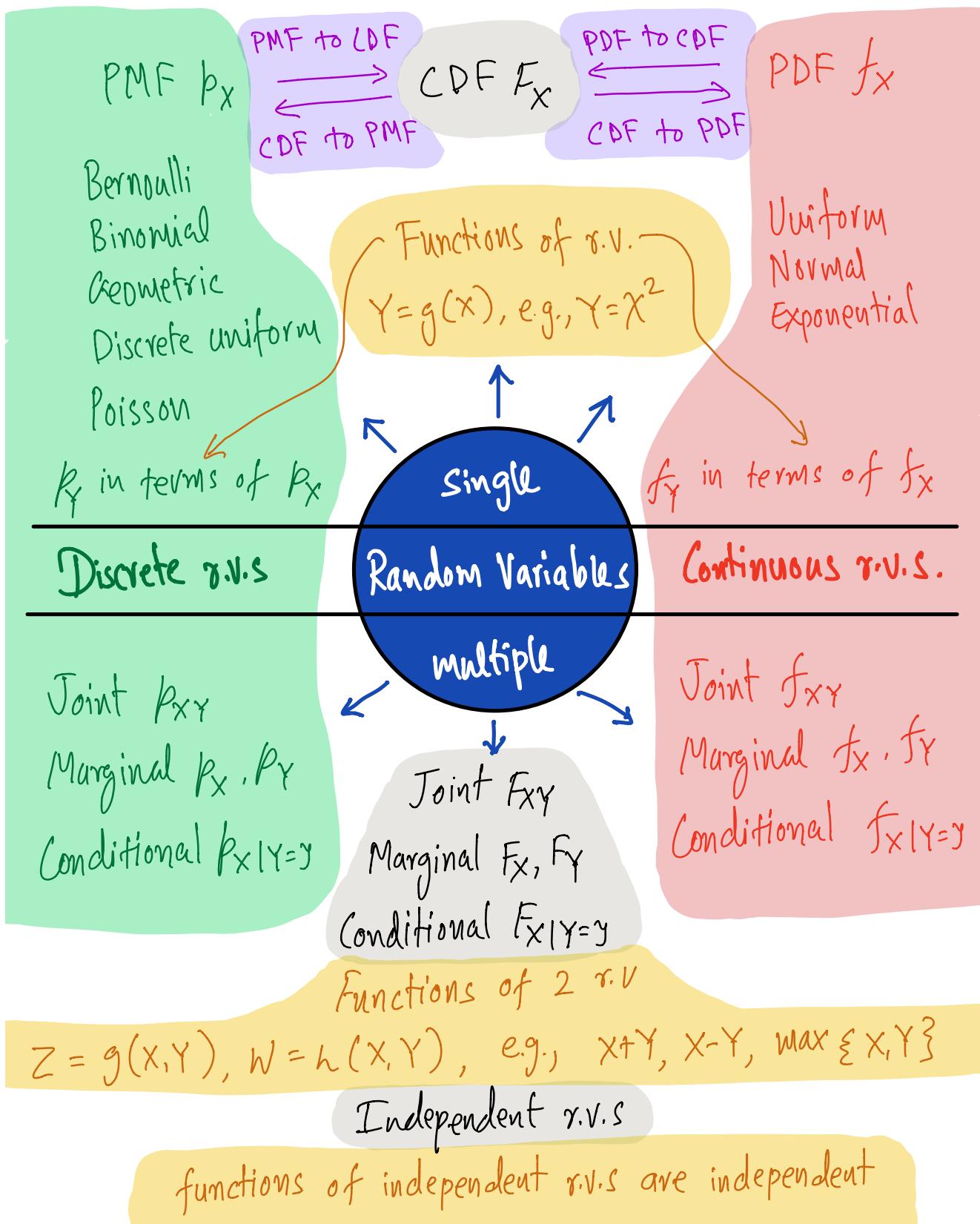
$$\begin{aligned} p\text{-value} &= P(\bar{X} \geq 69.3 \text{ when } \mu=68) \\ &\quad + P(\bar{X} \leq 66.7 \text{ when } \mu=68) \\ &= P(Y_0 \geq 69.3) + P(Y_0 \leq 66.7) \\ &= 2 P(Y_0 \leq 66.7) \quad (\because f_{Y_0} \text{ is symmetric around 68}) \\ &= 2 P(Z \leq -\frac{1.3}{.6}) \\ &= 2 \Phi_z(-2.16) \\ &= 2 \cdot 0.0154 = 0.0308. \end{aligned}$$

Hypothesis testing

- That is, there is 3.1% chance that we get the sample mean 69.8 or worst when $\mu=68$.
- Hence, it is more likely that $\mu \neq 68$ and so we should reject H_0 .
- In practice, H_0 is rejected if p-value < 0.05.
A typical practice but not always.







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

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E(XY) - E(X)E(Y)\end{aligned}$$

Mean μ

$$E(X) = \sum x p_X(x) = \int x f_X(x) dx$$

$$\begin{aligned}E(g(X)) &= \sum_x g(x) p_X(x) \\ &= \int g(x) f_X(x) dx\end{aligned}$$

$$E(X+Y) = E(X) + E(Y)$$

$$E(cX) = cE(X)$$

$$E(c) = c$$

Parameters & Properties

If X and Y are independent:

$$E(XY) = E(X) \cdot E(Y)$$

Markov inequality:

$$P(X \geq a) \leq \frac{E(X)}{a}$$

WLLN

For iid X_1, \dots, X_n with μ and σ^2

$$P(|\bar{X} - \mu| \geq \epsilon) \rightarrow 0 \text{ as } n \rightarrow \infty$$

Median: x such that
 $P(X \leq x) \geq .5$ and
 $P(X \geq x) \leq .5$

Standard deviation σ
 $SD(X) = \sqrt{\text{Var}(X)}$

Variance σ^2

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

$$\text{Var}(X) \geq 0$$

$$\text{Var}(c) = 0$$

$$\text{Var}(X+c) = \text{Var}(X)$$

$$\text{Var}(cX) = c^2 \text{Var}(X)$$

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

Chebyshhev inequality:

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

CLT

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

as $n \rightarrow \infty$