5CS037 – Concepts and Technology of AI.
Lecture – 03
{Discrete}Probability A Big Picture!!

About a Probability Theory for Machine Learning.

Siman Giri {Module Leader → 5CS037}

Disclaimer!!!
This Module Only discusses the Probability in Discrete Probability.

Learning Outcome:

- By the end of this week:
  - "Develop an intuitive understanding of probability theory and its core concepts, including axioms, random variables, and probability distributions, while appreciating their significance for understanding and applying machine learning principles."
  - We will discuss:
    - **Foundations of Probability:**
      - Understand **What**, **Why** and **How** of **Probability**, …
      - Various **Rules** and **Properties** of **Probability**, …
    - **Random variable and Probability distribution:**
      - **Understand** and **Analyze** the idea of **Probability distribution** and there **properties**.

1. What, Why and How of Probability?

{Probability – The Measure of Uncertainty?}

1.1 What is Probability?

- We are all familiar with the phrases:
  - "the **probability** that a coin will land heads is **0.5**".
  - "the **expected probability** of rolling a 5 on a **fair six sided die** is $\left\{\frac{1}{6} = 0.1667\right\}$."
- But what does this mean?
  - **Where do the numbers come from?**
    - or **How did we calculated the number?**

1.2 Where do Probability Comes From?

- We are all familiar with the phrases:
  - "the **probability** that a coin will land heads is **0.5**".
  - "the **expected probability** of rolling a 5 on a **fair six sided die** is $\left\{\frac{1}{6} = \mathbf{0.1667}\right\}$."
- But what does this mean?
  - Where do the numbers come from?
    - or How did we calculated the number?
- **Intuition:**
  - In some **process**, several outcomes are possible:
    - two outcomes when you flip a fair coin, $\{\mathbf{Head}, \mathbf{Tail}\}$ **.**
    - six outcomes when you roll a fair six sided die, $\{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}\}$**.**
  - When the process is repeated a large number of times, each outcome occurs with a characteristic relative frequency, or probability.
    - If a particular outcome happens more often than another outcome we say it is more probable outcome, we say it is more probable.
  - {The **process** are also called **Experiment** or **Random Experiment or Chance Experiment** in Particular.}

1.2.1 Probability can arise from →

- In *General* **Probability** can arise in **two contexts**:
  - **In actual repeated process or experiments**:
    - **For example:**
      - You record the color of **1000 cars** driving by.
        - **57** of them are **green**. You estimate the probability of a car being green as
          - $\frac{\mathbf{57}}{\mathbf{1000}}$ { **57 out of 1000**} → A relative frequency approach .
  - **In idealized conceptions of a repeated process or experiment**:
    - **Rolling a Die**:
      - Imagine rolling a fair six-sided die. Each side has an equal chance of showing up, so the probability of rolling a **5** is **1 out of 6** $= \frac{\mathbf{1}}{\mathbf{6}}$**.**
    - **Modeling Heights**:
      - To understand how people's heights are spread out, you can use a bell-shaped curve called the normal distribution. This curve helps estimate how common different heights are.
        - {From mathematical theory like law of large numbers and empirical evidence i.e. we performed the experiment large numbers of time and every time we got the same result – match our expectation.}

1.3 How to compute Probability?

- There are **three popular interpretation** of probability:
  - **Frequentist Interpretation (aka Empirical Interpretation):**
    - Interprets probability as the long-run frequency of an **event** occurring in repeated trials/experiment or process.
      - The probability of landing heads is 0.50 because, in the long run, if we flip the coin many times, half the flips are expected to result in heads.
    - Backed by **Law of Large Numbers**.
    - {This approach was further developed and formalized by **A. Kolmogorov** and popularly called **Axiomatic or Mathematical Interpretation**, this is what we will based our talk today and this week.}
  - **Bayesian Interpretation:**
    - Interprets probability as **a degree of belief**, updated using prior knowledge and new evidence.
      - The probability of landing heads is 0.5 because we assume prior belief that the coin is fair, and this belief can be updated with more evidence (e.g., results of coin flips).
    - Backed by **Bayes Rule**.
    - {We will discuss Bayes rule, when we introduce probabilistic approach of machine learning.}

Terminology Alert: Law of Large Numbers!!

- The **Law of Large Numbers (LLN)** is a fundamental theorem in probability theory that states:
  - *"As the number of trials or observations increases, the average of the observed outcomes will converge to the expected value (the true probability) of the event."*
- **Example:**
  - If you repeatedly flip a fair coin:
    - In a small number of flips, the proportion of heads may vary significantly
      - e.g., 3 heads out of 4 flips = 0.75.
  - However, as the number of flips increases, the proportion of heads will get closer to 0.50, the expected probability.
- The **LLN** ensures that observed probabilities **stabilize** with a large number of trials.

1.3.1 Compute Probability: Classical Approach.

- **classical approach to Probability**
  - when the outcomes in the sample space of a **chance experiment are equally likely**, the probability of an event E, denoted by P(E), is the ratio of number of outcomes in the sample space:
    - $\mathbf{P(E) = \frac{number\ of\ outcomes\ favourable\ to\ E}{number\ of\ outcomes\ in\ the\ sample\ space.}}$
  - As per the definition:
    - probability measures consist of **counting the number of events.**
- This approach may only valid till events in a sample case are equally likely.

1.1 Then, What is Probability?

- "*Probability theory is nothing but common sense reduced to calculation*."
  — *Pierre Laplace, 1812*
- **Towards Probability Theory …**
  - In general, **Probability** is an **estimate or quantification of uncertainty** attached to an **event** related to some **process**.
    - Where do the **Uncertainty** may arise from:
      - Noisy Measurements,
      - Variability between samples,
      - Finite size of datasets
    - Probability provides a **consistent framework** for the quantification and manipulation of **uncertainty**.
      - In order to **model the behavior** of a **process** based on observed or empirical outcomes and make inferences about **future events**, we will discuss a formal mathematical interpretation of **probability theory**, grounded in the foundational work of **Andrey Kolmogorov.**

1.4 Foundation of Probability Theory!!!

- **Axiomatic or Mathematical Interpretation:**
  - **Definition:**
    - The axiomatic interpretation, developed by **Andrey Kolmogorov**, treats probability as a **mathematical function** defined on a **set of outcomes** (a sample space).
    - It is based on a set of formal axioms (rules) that probability values must satisfy, such as **non-negativity**, **normalization** (the total probability of all possible outcomes equals 1), and **additivity**.

A. Kolmogorov

Towards Understanding Commonsense that is Probability.

{2. Elements of Probabilistic Space!!!}

2.1 A Probabilistic Space.

- A probability model is a **mathematical representation** of:
  - **a random phenomenon {experiment}**,
    - defined by its **sample space**,
      - **outcomes within the sample space**,
        - and **probability measure**
          - {defines how to assign probability to each event governed by Axioms of Probability}.
- Thus, any Probabilistic Space has following elements:

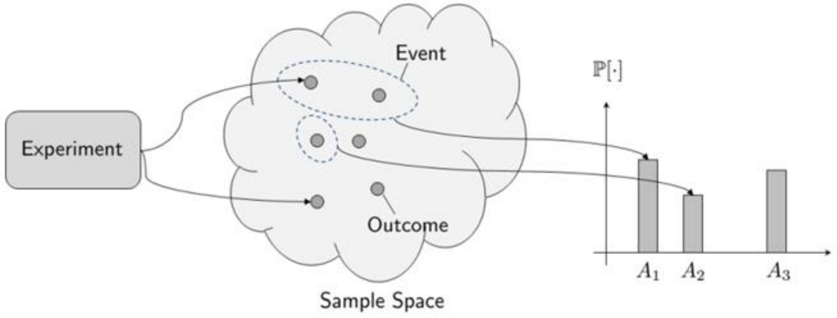


image by: S.Chan from Introduction to Probability for Datascience.

Fig: Elements of Probabilistic Space

2.2 {Chance} Experiment.

- **{Chance} Experiment aka Random Process:**
  - An chance experiment or random process is the process by which an observation is made {**Trial**}.
  - Key Features of Chance Experiment:
    - **Reproducibility**: The experiment can be conducted multiple times under identical conditions.
    - **Uncertainty**: The **exact outcome** of any single trial cannot be determined in **advance**.
    - **Sample Space**: The set of all **possible outcomes**, known as the sample space, is **well-defined**.

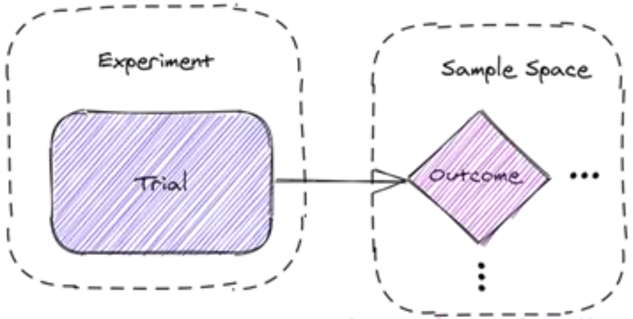


image by: Ash Bellet

Fig: A Chance Experiment and Trial.

2.3 Sample space.

- A **Sample Space** is the **set** of all possible outcomes of an experiment.
- A sample spaces may be **discrete** or **continuous.**
  - Continuous sample spaces are usually intervals of $\mathbb{R}$ or $\mathbb{R}^{n}$ used to model time, position, temperature, etc.
- Sample Space is denoted with Ω: {Read Omega} or simply capital letter "**S**“,
  - {cautions: we may use the sign interchangeably}
- For Example:
  - **Coin Flip: S = {Heads, Tails}(Discrete)**
  - **Flipping two coins: S = {(H,H),(H,T),(T,H),(T,T)}**
  - **YouTube hours in a day:** $S = \{x | x \in \mathbb{R}, 0 \leq x \leq 24\}$ **(Continuous)**

Disclaimer!!! This module we will only discuss about Discrete Probability Measures i.e. sample space composed of Discrete Measurements.

2.4 Events in a Event Space.

- An **event F** in a discrete sample **space S** is a collection of sample points(outcomes) i.e. any subset of **S or Ω**.
  - $\mathbf{F \subseteq S}$ {Usually denoted with F}
- **Simple Discrete Event:**
  - A simple event is an event consisting of exactly **one outcome**.
  - For example – For single coin flip $\mathbf{S \coloneqq \{H, T\}}$**:**
    - Event: $\mathbf{F_1 = \{Getting\ Head\} = \{H\}}$ → **This is a simple event**.
- **Event can also be Continuous:**
  - For example – Youtube Hours in a day $\mathbf{S = \{x | x \in \mathbb{R}, 0 \le x \le 24\}}$
    - Event: $\mathbf{F_{wasted\ day} = \{Youtube\ \ge 5\ hours\} = \{x | x \in \mathbb{R}, 5 \le x \le 24\}}$
- **Event can also be set of functions:**
  - For Example – Toss a fair coin an infinite number of times $S = \{f : \mathbb{N} \to \{H, T\}\}$
    - Event: $\mathbf{F \coloneqq The\ First\ two\ tosses\ are\ tails \coloneqq \{f \in S | f(1) = T\ and\ f(2) = T\}}$.
- In the world of probability: events are binary: they either happen or they do not.

Disclaimer!!! This module we will only discuss about Discrete Probability Measures i.e. sample space composed of Discrete Measurements.

2.4.1 Event Space in Probability.

- An **event space** is the collection of all possible events associated with a given sample space.
  - Events are subsets of the sample space, and the **event space** includes all such subsets that are **relevant** to the probability experiment.
- To Summarize:
  - **Sample Space (Ω or S):** The set of all possible outcomes of a random experiment.
  - **Event (F)**: A **subset** of the **sample space (F ⊆ S)** that corresponds to a particular outcome or group of outcomes.
  - **Event Space**($\mathcal{F}$){read sigma algebra}: A collection of subsets of the **sample space**, including:
    - the empty set is included – i.e. $\boldsymbol{\phi} \in \boldsymbol{\mathcal{F}}$ **;**
    - the entire sample space is included – **i.e.** $\mathbf{S} \in \boldsymbol{\mathcal{F}}$;
    - the complement of an event is also in the event space – i.e. If $\mathbf{A} \in \boldsymbol{\mathcal{F}}$ then $\mathbf{A^c} \in \boldsymbol{\mathcal{F}}$;
    - the event space is closed under countable unions – i.e. If $\mathbf{A_1}, \mathbf{A_2}, \ldots \in \boldsymbol{\mathcal{F}}$ then $\mathbf{A_i} \in \boldsymbol{\mathcal{F}}$
      - Example – **Event Space** – **Consider, Flipping a coin**. The event space is given as:
        - $\boldsymbol{\mathcal{F} = \{\phi, \{H\}, \{T\}, \Omega\}}$
- These properties ensure that probabilities can be consistently assigned to events.

2.4.2 Events that are Complex.

Complex Events

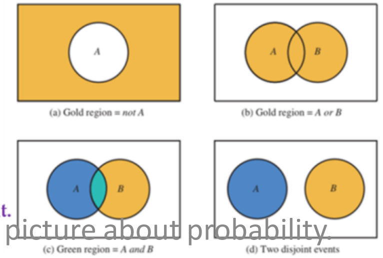
- A **complex event** refers to an event composed of
  - multiple simple events combined using logical operators such as
    - union (∪), intersection (∩), or complement (A’).
  - It may involve **more than** one event but does not
    - specifically require their **simultaneous occurrence.**
- **For example – Rolling a die:**
  - **Event A**: Rolling an even number:
    - $\mathbf{F_A \coloneqq \{2, 4, 6\}.}$
  - **Event B**: Rolling a number greater than 4:
    - $\mathbf{F_B \coloneqq \{5, 6\}.}$
  - **Complex Event**: Rolling an even number **or** a number greater than 4 :
    - $\mathbf{(F_A \cup F_B) = \{2, 4, 5, 6\}}$

Joint Events

- A **joint event** refers to the simultaneous occurrence of two or more events.
- It is specifically concerned with their **intersection** (A∩B) i.e., the outcomes common to both events.
- For example – Rolling a die:
  - **Event A**: Rolling an even number:
    - $\mathbf{F_A \coloneqq \{2, 4, 6\}.}$
  - **Event B**: Flipping heads on a single coin:
    - $\mathbf{F_B \coloneqq \{H\}.}$
  - **Joint Event**: Rolling an even number and Fliping a head.
    - $\mathbf{(F_A \cap F_B) = \{(2, H), (4, H), (6, H)\}.}$

In Figure →

a) not A $\{A^c\}$
b) A or B $\{A \cup B\}$
c) A and B $\{A \cap B\}$
d) Disjoint.



(a) Gold region = *not A*

(b) Gold region = *A or B*

(c) Green region = *A and B*

(d) Two disjoint events

image from internet : may subjected to copyright.

2.5 Probability Measure.

- A **probability measure** is a function that maps an events in an **event space** ($\mathcal{F}$) to a real number [0,1] satisfying the **axioms of probability**:
  - $\mathbb{P}: \mathcal{F} \rightarrow [\mathbf{0}, \mathbf{1}]$

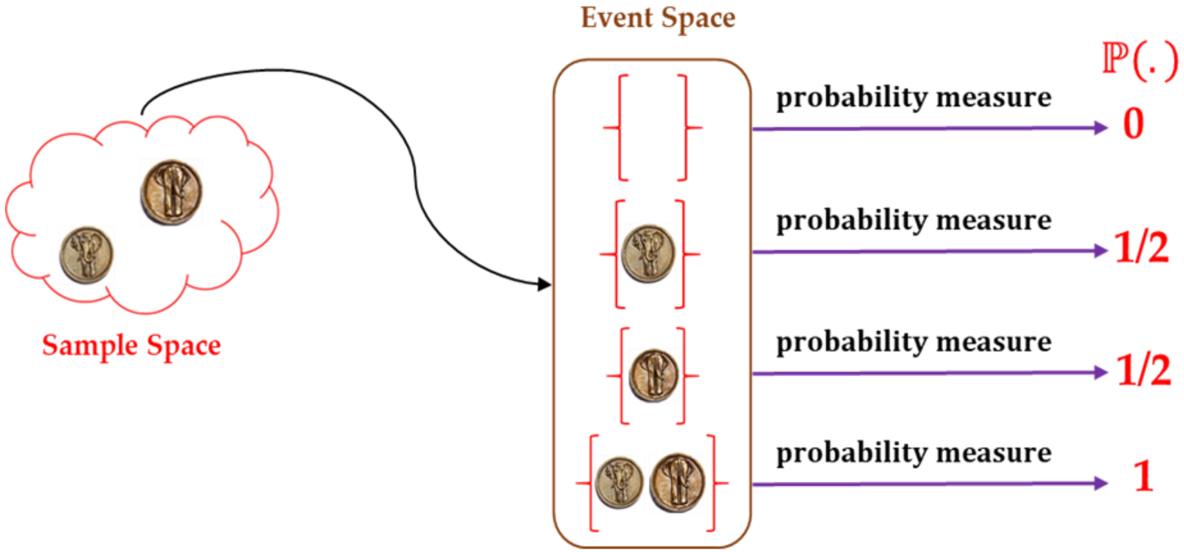


Fig: A Probabilistic Space.

←In a Figure:

Let $(\boldsymbol{S}, \boldsymbol{\mathcal{F}}, \mathbb{P})$ be a probability space, where

$\boldsymbol{S}$: Sample space (set of all possible outcomes).

$\boldsymbol{\mathcal{F}}$: Event space.

$\mathbb{P}$: $\boldsymbol{\mathcal{F}} \rightarrow [\mathbf{0}, \mathbf{1}]$: A probability measure.

2.6 Axioms of Probability.

- The **axioms of probability** are the foundational rules that any probability measure must satisfy to ensure consistency and validity in probability theory.
- These axioms, introduced by **Andrey Kolmogorov**
  - (that is why also known as Kolmogorov Axioms of Probability),
    - define how probabilities are assigned to events in a mathematically rigorous way.
- For a probability space defined by $(\boldsymbol{S}, \boldsymbol{\mathcal{F}}, \mathbb{P})$, where
  - $\boldsymbol{S}$: Sample space (set of all possible outcomes).
  - $\boldsymbol{\mathcal{F}}$: Event space.
  - P: $\boldsymbol{\mathcal{F}} \rightarrow [\mathbf{0}, \mathbf{1}]$: A probability measure.
- **Axioms are defined as →**

- **The Axioms of Probability are:**
  - **Non Negativity**: $\mathbf{P(A) \geq 0\ \forall A \in \mathcal{F}}$.
  - **Normalization**: $\mathrm{P(S) = 1}$.**{Sum of probability all possible outcomes}**
  - **Additivity (Countable Additivity):**
    - For any sequence of mutually exclusive events
    - $\mathbf{A_1, A_2, \ldots}$ $\left(\mathbf{i.e.\ A_i \cap A_j = \phi\ for\ i \neq j}\right)$:
      - $\mathbf{P(\cup_{i=1}^{\infty} A_i\ )) = \sum_{i=1}^{\infty} P(A_i)}$.
    - For example: For two mutually exclusive **events** A and B:
      - $\mathbf{P(A \cup B) = P(A) + P(B)}$.

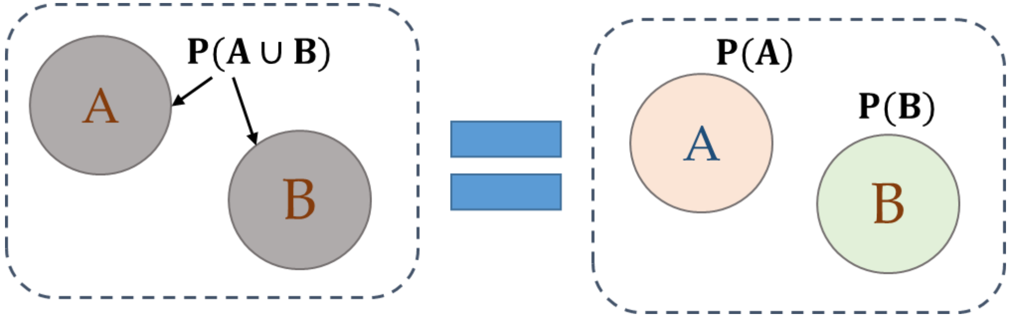
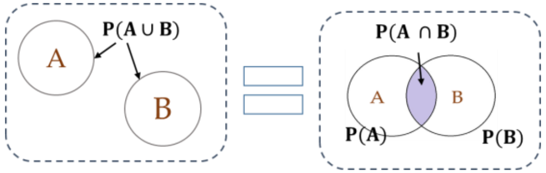


Fig: Understanding Additivity.

2.7 Corollaries derived from the axioms.

- Properties of Probability derived from Axioms of Probability:
- Corollary 1 – Probability of the Empty Set:
  - The probability of the empty set ∅ (an event with no outcomes) is zero i.e. $\mathbf{P(\phi) = 0.}$
- Corollary 2 – Probability of Complements:
  - The probability of the complements of an event $\mathbf{A}$ (denoted $\mathbf{A^c}$) is: $\mathbf{P(A^c) = 1 \; - P(A).}$
- Corollary 3 – Probability of Subsets:
  - If $\mathbf{A \subseteq B}$, then $\mathbf{P(A) \leq P(B).}$
- Corollary 4 – Bounds of Probability:
  - For any event A: $\mathbf{0 \leq P(A) \leq 1.}$
  - For any event A and B: $\mathbf{P(A \cup B) \leq P(A) + P(B)}$ {Union Bound}

- Corollary 5 – Inclusion – Exclusion or Probability of Union of events:
- For any events A and B: $\mathbf{P(A \cup B) = P(A) + P(B) - P(A \cap B).}$



- General Form or Compact representation of Inclusion – Exclusion:
  - $\mathbf{P(\cup_{i=1}^{n} A_i) = \sum_{\phi \neq J \subseteq \{1,2,\ldots,n\}} (-1)^{|J|+1} P(\cup_{j \in J} A_j);}$
  - **Where**
    - $\mathbf{J}$ is a subset of indices {1,2,…,n}
    - $\mathbf{|J|}$ is the size of the subset $\mathbf{J}$ (i.e., the number of events in the intersection).
    - The alternating sign $\mathbf{(-1)^{|J|+1}}$ adjusts for over-counting.

Assigning Probability to an Events {Discrete}.

{3. Based on Axiomatic Approach}

3.1 Frequentist Approach to Probability.

- **Frequentist approach:**
  - The **Frequentist approach** to probability defines the probability of an event as the long-run relative frequency with which the event occurs in repeated independent trials of an experiment.
    - Probability ➔ relative size of set of event w.r.t the sample space i.e. Out of total Outcomes how many are favorable outcomes i.e.
    - For any event $\mathbf{E} \in \mathcal{F}$.
      - $\mathbf{P(E)} = \frac{\text{\# Count number of favourable outocmes} \in \mathbf{E}}{\text{\# Count number of outcomes in the sample spce} \in \mathbf{\Omega}}$
- **Example:**
- What is the probability of the coin landing on heads?
  - **Sample space**: $\mathbf{S} \rightarrow \{\mathbf{H}, \mathbf{T}\}$
  - **Event:** $\{\mathbf{H}\} \in \mathcal{F} \rightarrow \{\boldsymbol{\phi}, \mathbf{H}, \mathbf{T}, \mathbf{S}\}$
  - $\mathbf{P(H)} = \frac{\text{\# Number of favourable Outcomes}}{\text{\# Number of possible Outocmes}} = \frac{1}{2}$

Cautions!! Slide is only valid for Discrete Sample spaces,
it is not possible to count an event in continuous sample space, thus probability are calculated relative to some function.

3.1.1 Key Features of Frequentist Approach.

- **Empirical Probability:**
  - The probability of an event is determined by conducting the experiment repeatedly and observing the proportion of times the event occurs.
- **Long-Run Frequency**:
  - As the number of trials increases, the observed frequency of the event converges to a fixed value, which is interpreted as the probability of the event.
- **Objective Interpretation**:
  - Probability is considered a property of the system or experiment, independent of personal belief or prior information.

- **Distinction from Classical Approach:**
  - The **classical approach** assigns probabilities based on equally likely outcomes and a known sample space.
  - The **Frequentist approach** estimates probabilities through repeated trials and empirical data, **without assuming** equally likely outcomes.
- In summary, the **Frequentist approach** treats probability as
  - a long-run frequency of an event occurring in repeated trials,
    - emphasizing empirical data and observations
  - rather than **theoretical assumptions** about the **nature of the experiment**.

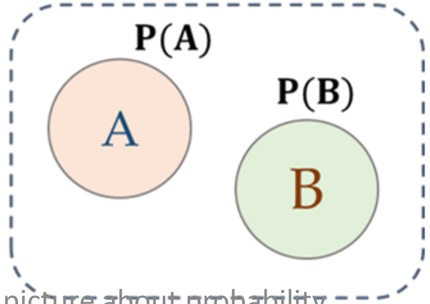
3.1.2 Revisiting – Axiomatic Approach to Probability.

- The **axiomatic approach** defines probability abstractly
  - using a set of axioms,
    - allowing for a more general and flexible interpretation of probability
      - that can be applied to both **theoretical(classical) and empirical (Frequentist) problems**.
- This approach provides a rigorous mathematical foundation for probability theory, making it applicable to a wide variety of situations, including those with complex or infinite sample spaces.

A. Kolmogorov

3.2 Probability for Mutually Exclusive Events.

- **Mutually exclusive events** are events that **cannot occur simultaneously**.
- In other words, the occurrence of one event precludes the occurrence of the other(s). For example:
  - In a single coin toss, the events
    - "**Heads**" and "**Tails**" are mutually exclusive.
  - In rolling a die, the events
    - "**rolling a 3**" and "**rolling a 4**" are mutually exclusive.
- **Definition:**
  - Two events A and B are mutually exclusive if:
      - $\mathbf{A \cap B = \phi}$;
    - where $\mathbf{A \cap B}$ is the intersection of A and B, representing outcomes common to both. Since $\mathbf{A \cap B = \phi}$,
    - the probability of both events occurring together is:
      - $\mathbf{P(A \cap \mathit{B}) = 0}$.



← Fig: Mutually Exclusive Events.

3.2.1 Probability for Mutually Exclusive Events.

- **Assigning Probability to Mutually Exclusive Events:**
  - When events are mutually exclusive, the probability of their union is:
  - $\mathbf{P(A \cup B) = P(A) + P(B).}$
  - {Based on additivity axiom of probability and also known as Union or Additive law of probability}.
- If we are not sure about the exclusivity of two events A and B , based on inclusion – exclusion we write:
  - $\mathbf{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$
- **Example – Tossing a Coin:**
  - Suppose you toss a fair coin. What is the Probability of getting head or Tail. Let:
    - **A={Heads}, B={Tails}**
  - Since A and B are mutually exclusive:
    - $\mathbf{P(A) = \frac{1}{2}; P(B) = \frac{1}{2}}$
  - The probability of either heads or tails is:
    - $\mathbf{P(A \cup B) = P(A) + P(B) = \frac{1}{2} + \frac{1}{2} = 1.}$

3.2.2 Probability for Mutually Exclusive Events.

- **Assigning Probability to Mutually Exclusive Events:**
  - When events are mutually exclusive, the probability of their union is:
  - $\mathbf{P(A \cup B) = P(A) + P(B)}$.
  - {Based on additivity axiom of probability and also known as Union or Additive law of probability}.
- If we are not sure about the exclusivity of two events A and B , based on inclusion – exclusion we write:
  - $\mathbf{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$
- **Example – Tossing a Coin:**
  - Suppose you toss a fair coin. What is the Probability of getting head or Tail. Let:
    - **A={Heads}, B={Tails}**
  - Since A and B are mutually exclusive:
    - $\mathbf{P(A) = \frac{1}{2}; P(B) = \frac{1}{2}}$
  - The probability of either heads or tails is:
    - $\mathbf{P(A \cup B) = P(A) + P(B) = \frac{1}{2} + \frac{1}{2} = 1}$.

Q: Can not two events be related, If yes then how do you calculate the Probability (for joint events)?

3.3 Probability for with and without replacement.

- Example-with replacement:
  - In a bag, there are 5 red balls and 3 blue balls. A ball is drawn from the bag, its color is noted, and then **it's placed back in the bag**. This process is repeated three times. What is the probability of drawing exactly 2 red balls and 1 blue ball?
  - Solution:

- We have – for individual outcome – Probability is :
  - Probability for Red Ball - R: $\mathrm{P(R)} = \frac{5}{8};$
  - Probability for Blue Ball - B: $\mathbf{P(B)} = \frac{3}{8};$
- Our event of interest is:
  - We want to calculate the probability of drawing exactly 2 red balls and 1 blue ball in 3 draws. There are three ways this can happen:
    $\mathbf{Events} \coloneqq \{\mathbf{Draw\ 2\ red\ balls\ and\ 1\ blue\ ball}\} \coloneqq \{\mathbf{RRB, RBR, BRR}\}\{\#\ \mathbf{note\ order\ does\ not\ matter}\}$
  - Let's Calculate the Probability:

$$\mathrm{P(RRB)} = \frac{5}{8}.\frac{5}{8}.\frac{3}{8}$$
$$\mathrm{P(RBR)} = \frac{5}{8}.\frac{3}{8}.\frac{5}{8}$$
$$\mathrm{P(BRR)} = \frac{3}{8}.\frac{5}{8}.\frac{5}{8}$$

Thus, total Probability: $\mathbf{Total\ probability} = \mathbf{P(RRB)} + \mathbf{P(RBR)} + \mathbf{P(BRR)}.$

- Example-without replacement:
  - In a bag, there are 5 red balls and 3 blue balls. A ball is drawn from the bag, its color is noted, and then **it's not placed back in the bag**. This process is repeated three times. What is the probability of drawing exactly 2 red balls and 1 blue ball?
  - Solution:

- We have – for individual outcome – Probability is :
  - Probability for Red Ball - R: $\mathrm{P(R)} = \frac{5}{8};$
  - Probability for Blue Ball - B: $\mathbf{P(B)} = \frac{3}{8};$
- Our event of interest is:
  - We want to calculate the probability of drawing exactly 2 red balls and 1 blue ball in 3 draws. There are three ways this can happen:
    $\mathbf{Events} \coloneqq \{\mathbf{Draw\ 2\ red\ balls\ and\ 1\ blue\ ball}\} \coloneqq \{\mathbf{RRB, RBR, BRR}\}\{\#\ \mathbf{note\ order\ does\ not\ matter}\}$
  - Let's Calculate the Probability:

$$\mathrm{P(RRB)} = \frac{5}{8}.\frac{4}{7}.\frac{3}{6}$$
$$\mathrm{P(RBR)} = \frac{5}{8}.\frac{3}{7}.\frac{4}{6}$$
$$\mathrm{P(BRR)} = \frac{3}{8}.\frac{5}{7}.\frac{4}{6}$$

Thus, total Probability: $\mathbf{Total\ probability} = \mathbf{P(RRB)} + \mathbf{P(RBR)} + \mathbf{P(BRR)}.$

Q: What did you Observe?

3.3 Probability for with and without replacement.

- Example-with replacement:
  - In a bag, there are 5 red balls and 3 blue balls. A ball is drawn from the bag, its color is noted, and then **it's placed back in the bag**. This process is repeated three times. What is the probability of drawing exactly 2 red balls and 1 blue ball?
  - Solution:

- We have – for individual outcome – Probability is :
  - Probability for Red Ball - R: $P(R) = \frac{5}{8};$
  - Probability for **Blue Ball - B:** $\mathbf{P(B)} = \frac{3}{8};$
- Our event of interest is:
  - We want to calculate the probability of drawing exactly 2 red balls and 1 blue ball in 3 draws. There are three ways this can happen:
    Events ≔ {Draw 2 red balls and 1 blue ball} ≔ {RRB, RBR, BRR}{# note order does not matter}
  - Let's Calculate the Probability:

$$P(RRB) = \frac{5}{8}.\frac{5}{8}.\frac{3}{8}$$
$$P(RBR) = \frac{5}{8}.\frac{3}{8}.\frac{5}{8}$$
$$P(BRR) = \frac{3}{8}.\frac{5}{8}.\frac{5}{8}$$

Thus, total Probability: **Total probability = P(RRB) + P(RBR) + P(BRR).**

- Example-without replacement:
  - In a bag, there are 5 red balls and 3 blue balls. A ball is drawn from the bag, its color is noted, and then **it's not placed back in the bag**. This process is repeated three times. What is the probability of drawing exactly 2 red balls and 1 blue ball?
  - Solution:

- We have – for individual outcome – Probability is :
  - Probability for Red Ball - R: $P(R) = \frac{5}{8};$
  - Probability for **Blue Ball - B:** $\mathbf{P(B)} = \frac{3}{8};$
- Our event of interest is:
  - We want to calculate the probability of drawing exactly 2 red balls and 1 blue ball in 3 draws. There are three ways this can happen:
    Events ≔ {Draw 2 red balls and 1 blue ball} ≔ {RRB, RBR, BRR}{# note order does not matter}
  - Let's Calculate the Probability:

$$P(RRB) = \frac{5}{8}.\frac{4}{7}.\frac{3}{6}$$
$$P(RBR) = \frac{5}{8}.\frac{3}{7}.\frac{4}{6}$$
$$P(BRR) = \frac{3}{8}.\frac{5}{7}.\frac{4}{6}$$

Thus, total Probability: **Total probability = P(RRB) + P(RBR) + P(BRR).**

Q: What did you Observe?

3.4 Assigning Probability to an Events {Discrete}: Summary.

- To sum up – "Count" the probability of individual outcome, and sum them up for the probability of a collection of outcomes (= event).
- Thus we can say that:
  - To evaluate the probability of discrete events, the calculation depends on counting, which in turn depends on the **type of events**: {whether they are disjoint or joint}.
    - **Disjoint events** are always mutually exclusive, meaning they cannot occur simultaneously.
      - $\mathbf{P(A \cup B) = P(A) + P(B).}$
    - **Joint events** can either be **independent (unrelated)** or **dependent (related or conditioned),** depending on whether the occurrence of one event **affects** the probability of the other.
      - $\mathbf{P(A \cap B) = \dfrac{\#\ Count\ outcomes\ of\ A\ and\ B.}{\#Total\ Outcomes\ in\ the\ sample\ space.}}$

3.5 Why Counting Matters?

- Counting is essential in probability because it provides the foundation for calculating probabilities, particularly for **discrete sample spaces**.
  - Counting helps distinguish between **disjoint (mutually exclusive)** and **joint (overlapping)** events:
    - For disjoint events,
      - **$P(A\cup B) = P(A)+P(B)$,**
        - and counting is straightforward since there's no overlap.
    - For joint events,
      - **$P(A\cup B) = P(A) + P(B) - P(A\cap B)$**
        - {Cautions !! This does not tell you whether relationship exist or not we discuss that in next section},
      - and counting must account for the overlap to avoid double-counting.
  - When events involve multiple stages or combinations (Complex events), counting becomes crucial:

- **For example – Solve it:**
  - We have a committee of n = 10 people and we want to choose a **chairperson**, **a** **vice-chairperson** and a **treasurer**. Suppose that 6 of the members of the committee are male and 4 of the members are female. What is the probability that the three executives selected are all female?
    - Solutions:
      - Favorable events: {C, VC, T: Male.}
      - Sample space: {(M1,M2,M3),(M2, M1,M3),……, (F3,F4,F1)}
      - P = ?
- Solving for probability requires counting of an total events in sample space,
  - which can be little complicated depending on the process/experiment and probability desired.
- Hope you know how to count, If **don't do not miss your tutorial, we will discuss on counting principle** – Hint **Permutations and Combinations !!!!**

Events that are Joint!!!

{4. Conditional Probability and Examination of Independence.}

4.1 Conditional Probability: Introduction.

- **Theory of conditional probability** provides the way to measure the probability when two events are dependent with each other, i.e. one event can only occur if another event has already happened.
- For **any event A and B**, if event **A is dependent on B** (or also called **conditioned on B**) and we know $\mathbf{P(B) \neq 0}$ Then **Conditional Probability of A given B** is :
  - $\mathbf{P(A|B) = \frac{P(A \cap B)}{P(B)}}$ is called conditional probability of A conditioned on B.
- The difference between conditional Probability $\{\mathbf{P(A|B)}\}$ and joint Event $\{\boldsymbol{P(A \cap B)}\}$ is the denominator they carry i.e.
  - $\mathbf{P(A|B) = \frac{P(A \cap B)}{P(B)}}$; Focuses on the likelihood of A given B has occurred.
    - **Conditional Probability** isolates the impact of B on A, focusing only on cases where B occurs.
  - $\mathbf{P(A \cap B) = \frac{P(A \cap B)}{P(\Omega)} = P(A \cap \mathit{B})}$; Represents the probability that both A and B occur Simultaneously.
    - **Joint Probability** measures the overlap of A and B in the entire sample space.

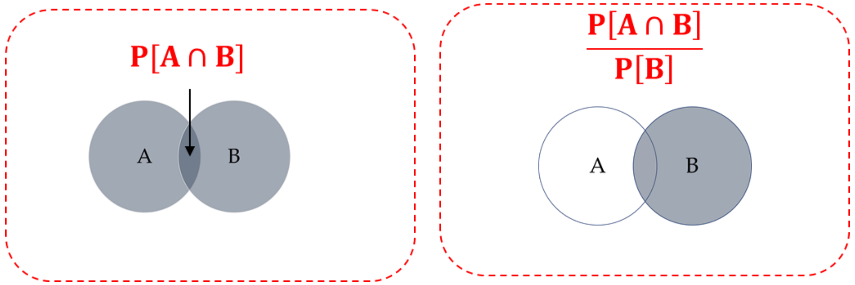


Fig: Illustration of Conditional Probability and comparison to Joint Probability.

4.2 Axioms of Conditional Probability.

Axioms of Probability.	Axioms – Conditional Probability.	
For any Event E – $P(E)$	For any Event E and F conditioned on – $P(E\|F)$	
$0 \le P(E) \le 1$	$0 \le P(E\|F) \le 1$	Non – negativity.
$P(S) = 1$	$P(S\|F) = 1$	Normalizations.
$P(E \text{ or } F) = P(E) + P(F)$	$P(E \text{ or } G\|F) = P(E\|F) + P(F\|G)$	Additive Probability.
$P(E^{c}) = 1 - P(E)$	$P(E^{c}\|F) = 1 - P(E\|F)$	Complement Probbaility.

Table: The Paradigm of Conditional Probability.

4.3 Example of Conditional Probability.

- Imagine Daraz wants to figure out the probability that a user will buy a smart watch (let's say W), based on knowing that they bought smart phone (let's say P).
- Solution:
  - We want : $\mathbf{P(Buy\ a\ smart\ watch|They\ bought\ a\ smart\ Phone) \Rightarrow P(W|P) = ?.}$
    - To answer, lets start with simpler question, what is the probability that a user bought a smart watch (W):
      - $\mathbf{P(W) = \frac{\#\ people\ who\ bought\ smart\ watch}{\#people\ on\ Daraz}}$
    - Now for second question; What is the probability that a user will buy smart watch given they bought smart phone.
      - $\mathbf{P(W|P) = \frac{P(W\ and\ P)}{P(P)} \approx \frac{\#\ who\ bought\ W\ and\ P/\#people\ in\ daraz}{\#who\ bought\ P/\#people\ in\ Daraz}}$;{Definition of Prob.}
      - $\mathbf{P(W|P) \approx \#\ of\ people\ who\ bought\ E\ and\ F/\ \#of\ people\ whou\ bought\ P.}$
  - Interpretation of the Result: Two Possibilities:
    - High Probability: Recommend a smart watch who bought phone.
    - Low Probability: Do not Recommend a smart watch who bought phone.

4.4 Independent Events: Introduction.

- We have introduced the conditional probability for event A conditioned on event B $\{\mathbf{P(A|B)}\}$ to capture the **partial information that event B provides about event A**.
  - An interesting and important special case arises when the occurrence of **B** provides **no information** and does not alter the probability that **A** will occur, i.e. when two events are independent:
  - **Two events** are said to be **independent** if **knowing** the **outcome** of **one event** does not **change your belief** about **whether or not** the **other event** will occur. Mathematically:
    - $\mathbf{P(A \cap B) = P(A) \times P(B)}$ {aka multiplicative rule of probability}
- **Example Scenario – Coin Tosses:**
  - Suppose you toss a coin twice:
    - **A = {Heads on the first toss},**
    - **B = {Heads on the second toss}.**
  - What is the probability of getting A and B:
    - Since the outcome of the first toss does not affect the second:
      - $\mathbf{P(A \cap B) = P(A) \times P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}}$.

4.5 Common Misunderstanding about Independence.

- **Independence is not same as Mutual Exclusivity:**
  - **Independent events:** Can occur simultaneously. For example, flipping a coin and rolling a die.
  - **Mutually exclusive events:** Cannot occur at the same time. For example, rolling a die and getting both a "3" and a "5."
- **Independence is not about disjointness:**
  - **Disjoint events** have no overlap, so **P(A ∩ B) = 0** . They cannot be independent unless one or both events have **P(A) = 0 or P(B) = 0**.
  - **Independent events** can overlap.
- **Independence is not about equal probabilities:**
  - Independent events do not require equal probabilities.
  - For example: rolling a die and flipping a coin are independent, even though
    - $\mathbf{P(Die\ Shows\ 4)} = \frac{1}{6}$ and $\mathbf{P(Coin\ shows\ head)} = \frac{1}{2}$.

4.6 Conditional Interpretation of Independence.

- Conditional Probability Definition:
  - The conditional probability of A given B is defined as:
    - $\mathbf{P(A|B) = \frac{P(A \cap B)}{P(B)}, provided\ P(B) > 0}$.
    - If A and B are independent ,
      - then the occurrence of B does not change the probability of A. i.e. $\mathbf{P(A|B) = P(A).}$
    - Substituting this into the definition of conditional probability:
      - $\mathbf{P(A) = \frac{P(A \cap B)}{P(B)}}$.
    - Multiplying through by $\mathbf{P(B)}$:
      - $\mathbf{P(A \cap B) = P(A) \cdot P(B).}$
  - This is the mathematical definition of independence.
- Iterative Extension to Multiple or n – Events:
  - The independence principles extends to more than two events. For n events $\mathbf{A_1, A_2, \ldots, A_n}$ that are mutually independent of one another – the independence equation also holds for all subsets of the events:
    - $\mathbf{P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) \cdot P(A_2) \cdots P(A_n) = \prod_{i=1}^{n} P(A_i)}$

4.7 Independence Example Exercise:

- Consider an experiment involving two successive rolls of a 4-sided die in which all 16 possible outcomes are equally likely and have probability 1/16.
  - (a) Are the events A = {1st roll is a 1}, B = {sum of the two rolls is a 5}, independent?
- Solution:

4.7.1 Independence Example Exercise – Solution.

- The sample space is:
  - $\mathbf{S = \{(i, j) : i, j \in \{1, 2, 3, 4\}\}}$
    - Where **i** is the result of the first roll A
    - and **j** is the result of the second roll B.
    - All 16 outcomes are equally likely with $\mathbf{P(\{A_i, B_j\}) = \frac{1}{16}}$.

(a) Are the events independent:

- Define the events:
  - $\mathbf{A = \{(1, j) : j \in \{1, 2, 3, 4\}\}}$,
    - $\mathbf{i.e.\ the\ first\ roll\ is\ 1 = \{(1, 1), (1, 2), (1, 3), (1, 4)\}}$
  - $\mathbf{B = \{(1, j) : i + j = 5\}}$,
    - $\mathbf{i.e.\ sum\ of\ the\ two\ rolls\ is\ 5\ = \{(1, 4), (2, 3), (3, 2), (4, 1)\}}$.
  - $\mathbf{A \cap B = \{(1, 4)\}\ \#\ common\ events\ in\ A\ and\ B}$

- Compute Probabilities:
  - Probability of A: $\mathbf{P(A) = \frac{4}{16} = \frac{1}{4}}$.
  - Probability of B: $\mathbf{P(B) = \frac{4}{16} = \frac{1}{4}}$.
  - Probability of $\mathbf{A \cap B}$: $\mathbf{P(A \cap B) = \frac{1}{16}}$.
- Check for Independence:
  - $\mathbf{P(A \cap B) = P(A) \cdot P(B)}$
    - $\mathbf{\frac{1}{16} = \frac{1}{4} \cdot \frac{1}{4}}$
      - $\mathbf{\frac{1}{16} = \frac{1}{16}}$.

4.8 Conditional Independence.

- **Conditional independence** extends concept of independence to three or more events.
  - We say that event **X and Y are conditionally independent given a third event Z** if the knowledge of **Z** makes the knowledge of **Y irrelevant to X**, and vice versa.
    - Formally, we write: **X ⊥ Y | Z**
      - read: **X** and **Y** are **conditionally independent** given **Z**. In this case, the conditional probability of **X** and **Y** together, given **Z**, factorizes as follows:
        - $\mathbf{P(x \cap y|z) = P(x|z)P(y|z) - -\{1\}}$
  - This is the key definition of **conditional independence**: after conditioning on **Z**, **the events X and Y become independent**.
    - In other words, knowing the value of **Y** provides **no additional information** about **X**, **once we know Z.**
  - The equation {1} can also be written as:
    - $\mathbf{P(x|y,z) = P(x|z)}$ **or equivalently** $\mathbf{P(y|x,z) = P(y|z)}$
- This means that, once we know **Z**, the probability of **X** does not depend on **Y**, and the probability of **Y** does not depend on **X**.

4.9 To sum up – joint Probability.

- Joint probability is the probability of **two or more events** happening at the **same time**. If we have two events **A and B**, the joint probability
  - $\mathbf{P(A \cap B)}$ (or sometimes written as $\mathbf{P(A, B)}$)
    - represents the likelihood of both **A and B** occurring together.
- For example, if we consider rolling two dice, let event A be {A: "rolling a 3 on the first die," } and event B be {B: "rolling a 5 on the second die.“}
  - The joint probability $\mathbf{P(A \cap B)}$ would be the probability of:
    - $\mathbf{P(A \cap B)} =$ **P(rolling a 3 on the first die) and P(5 on the second die).**
- For **independent events**, where one event does not influence the outcome of the other, the joint probability is simply the product of their individual probabilities:
  - $\mathbf{P(A \cap B) = P(A) \cdot P(B)}$
- For **dependent events**, where the occurrence of one event affects the probability of the other, the joint probability is found by multiplying the probability of one event by the conditional probability of the second event given the first:
  - $\mathbf{P(A \cap B) = P(A) \cdot P(B|A)}$
- In general**, joint probability distributions** describe **the probabilities** of **all combinations** of values for **multiple random variables**, which is essential for **understanding relationships** between them.

4.10 Chain Rule of Probability.

- The chain rule of probability express the probability of the intersection of multiple events in terms of conditional probabilities. For n events $\mathbf{A_1, A_2, \cdots, A_n}$ the chain rule is:
  - $\mathbf{P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})}$
- This rule allows us to compute the joint probability of multiple events step by step, starting from the probability of the first event and progressively conditioning on previous events.
- **Independent Events in the Chain Rule:**
  - If the n events $\mathbf{A_1, A_2, \cdots, A_n}$ are independent then, chain rule simplifies to
    - $\mathbf{P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) \cdot P(A_2) \cdot\cdot \cdots P(A_n) = \prod_{i=1}^{n} P(A_i)}$ {**result from slide 37.**}

Probability for Machine Learning.

{5. How do we Connect Sample Space and Events to data?.}

5.1 Random Variable.

- In a layman terms, random variables is a numeric {$\mathbb{R}$} value assigned to each outcome of a chance/random experiment.
  - In a way it can be thought of as a variable from programming, but this is a function that maps a uncertainty of any event in sample space with uncertainty.
  - For each element (outcomes) in a sample space, the random variable can take on exactly one value i.e. random variable must be measurable.
- A formal mathematical formulation or definition be:
  - A random variable **X** is a function **X**: $\boldsymbol{\Omega} \rightarrow \mathbb{R}$ that maps an outcome $\boldsymbol{\xi} \in \boldsymbol{\Omega}$ to a number **X**($\boldsymbol{\xi}$) on the **real line**.

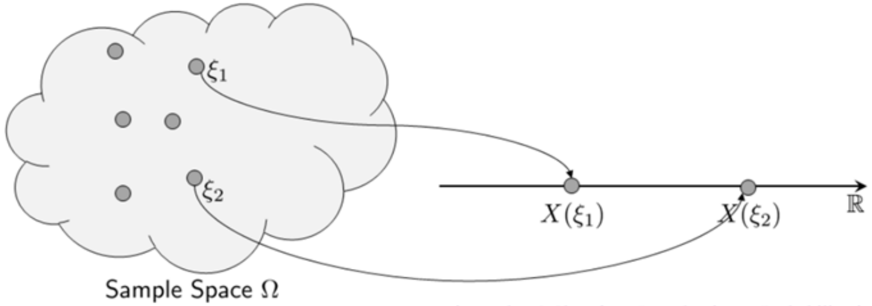


image by: S.Chan from Introduction to Probability for Datascience.

Fig: A definition of Random Variable.

5.2 Random Variables: Types.

- **Discrete Random Variable:**
  - A random variable Y is said to be discrete if it can assume only a finite or countably infinite number of distinct values.
  - A random variable is discrete if its set of possible values is a collection of isolated points on the number line.

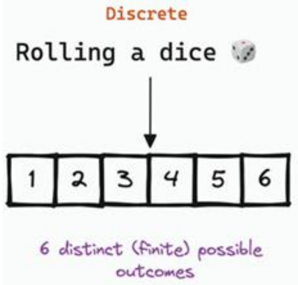


image from internet : may subjected to copyright.

- **Continuous Random Variable:**
  - A random variable Y is said to be Continuous if it can assume any value defined in-between interval or range.
  - The variable is continuous if its set of possible values includes an entire interval on the number line.

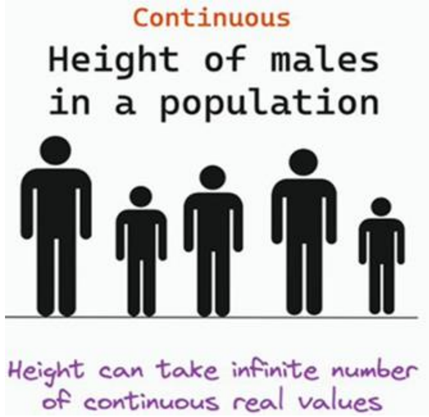
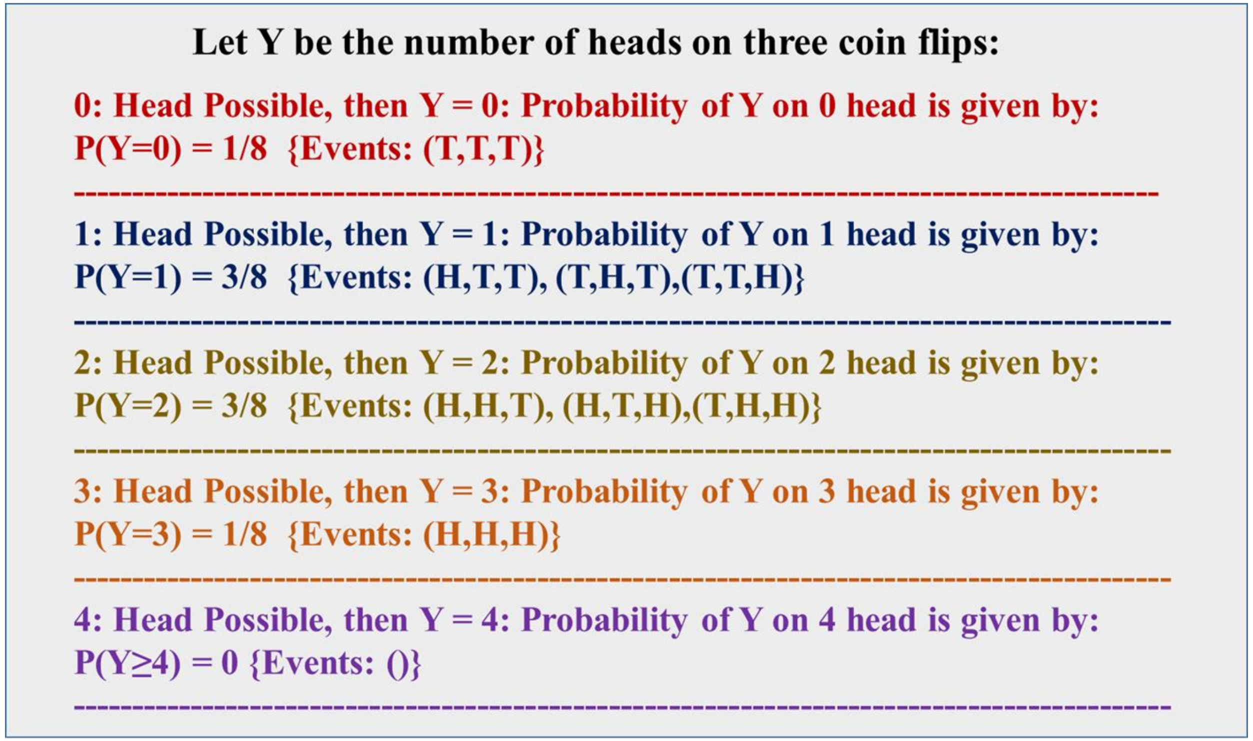

Continuous
Height of males in a population
Height can take infinite number of continuous real values

image from internet : may subjected to copyright.

5.3 Random Variable in Practice.

- **Experiment: Flip a three fair coins.**
- Observations of interest{**Y**}: Number of **"heads"** on the **three coins**.
  - **Y** is an **Random Variable**, which **maps observation of interest** to **real numbers**.
    - **A very big Question – What numbers and How?**
      - { By definition I **can map to any number in real number line**, but that does not mean we start mapping randomly by assigning any numbers,}
      - In practice we follow the convention: i.e.
      - Which in general is answer to the question:
        - **How many possible combinations of outcome for our observation of interest is possible?**
      - For above experiment → **how many possible outcomes are possible**?
- Let us find out:

5.3.1 Random Variable in Practice.



Let Y be the number of heads on three coin flips:

0: Head Possible, then Y = 0: Probability of Y on 0 head is given by:
P(Y=0) = 1/8 {Events: (T,T,T)}

1: Head Possible, then Y = 1: Probability of Y on 1 head is given by:
P(Y=1) = 3/8 {Events: (H,T,T), (T,H,T),(T,T,H)}

2: Head Possible, then Y = 2: Probability of Y on 2 head is given by:
P(Y=2) = 3/8 {Events: (H,H,T), (H,T,H),(T,H,H)}

3: Head Possible, then Y = 3: Probability of Y on 3 head is given by:
P(Y=3) = 1/8 {Events: (H,H,H)}

4: Head Possible, then Y = 4: Probability of Y on 4 head is given by:
P(Y≥4) = 0 {Events: ()}

Thus for **Our Random Variable Y** we can map to:

- $\mathbf{Y \rightarrow \{0, 1, 2, 3\}}$

To generalize we write:

- $\mathbf{(Y = y) : y \in \{0, 1, 2, 3\}}$

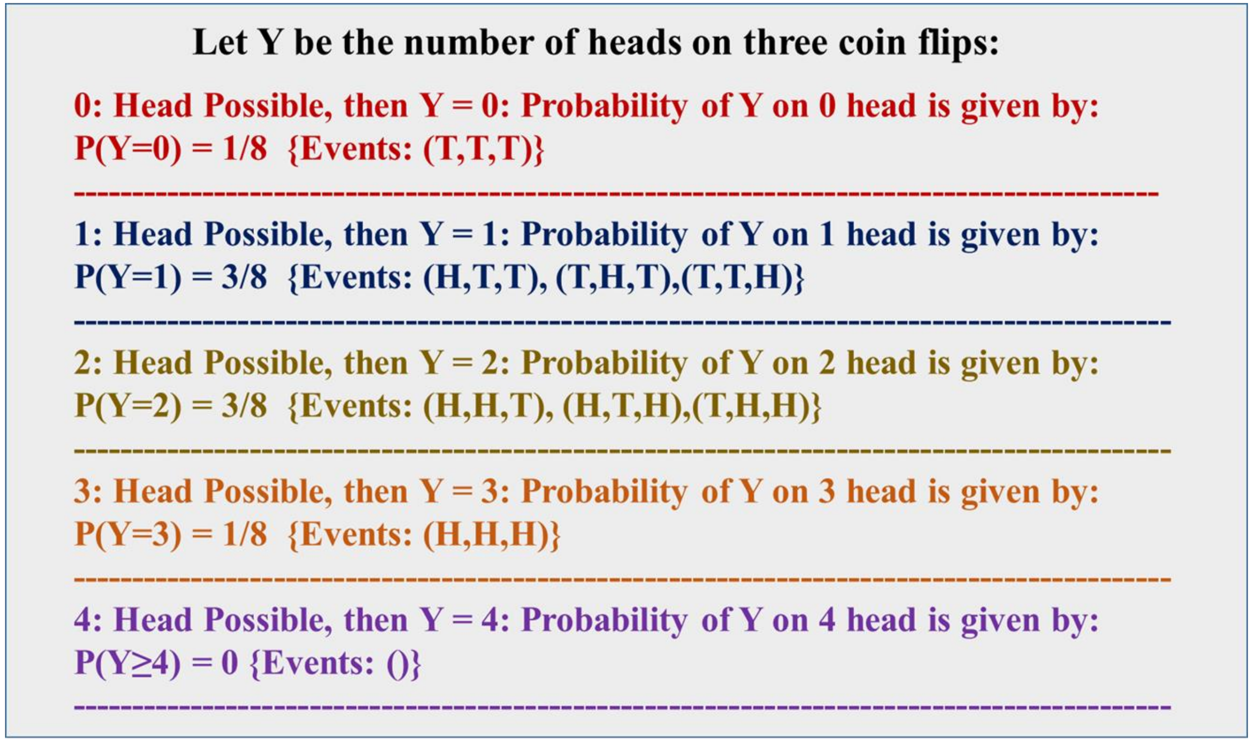
We assign a Probability as:

- $\mathbf{P(Y = y)}$
- $\mathbf{i.e.\ P(Y = 0), P(Y = 1), P(Y = 2) P(Y = 3)}$

Observation:

In this case a **Single Random Variable** $\mathbf{(Y = y)}$ models three different uncertainty values.

5.3.2 Random Variable in Practice.



Let Y be the number of heads on three coin flips:

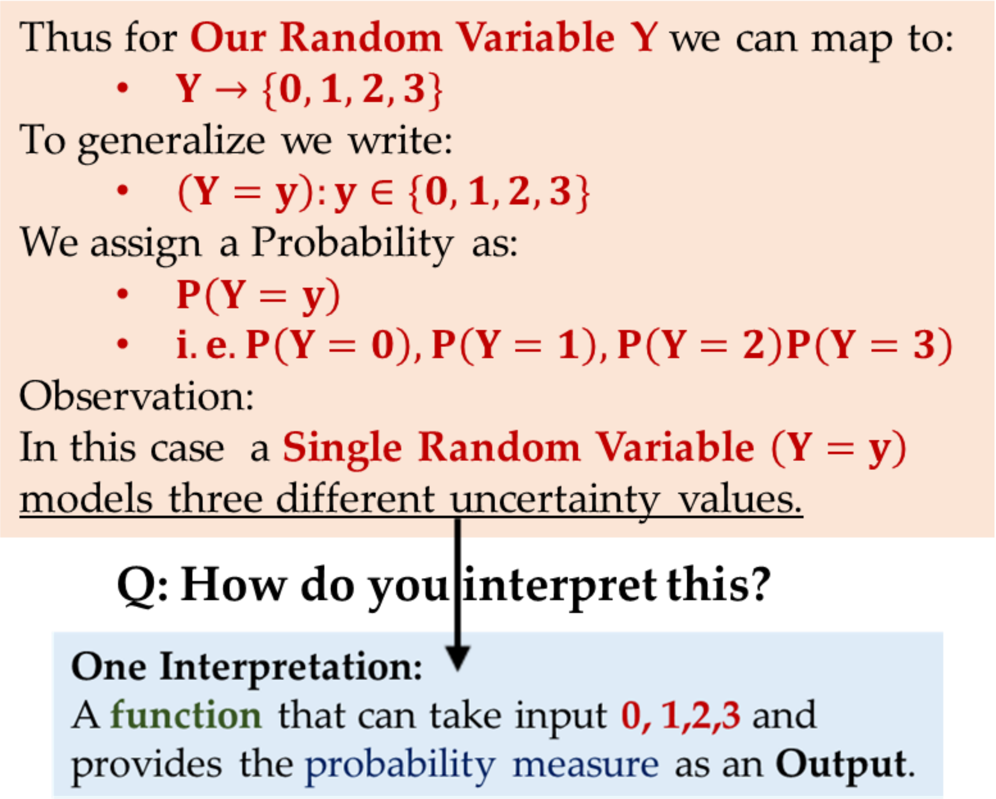
0: Head Possible, then Y = 0: Probability of Y on 0 head is given by:
P(Y=0) = 1/8 {Events: (T,T,T)}

1: Head Possible, then Y = 1: Probability of Y on 1 head is given by:
P(Y=1) = 3/8 {Events: (H,T,T), (T,H,T),(T,T,H)}

2: Head Possible, then Y = 2: Probability of Y on 2 head is given by:
P(Y=2) = 3/8 {Events: (H,H,T), (H,T,H),(T,H,H)}

3: Head Possible, then Y = 3: Probability of Y on 3 head is given by:
P(Y=3) = 1/8 {Events: (H,H,H)}

4: Head Possible, then Y = 4: Probability of Y on 4 head is given by:
P(Y≥4) = 0 {Events: ()}



Thus for **Our Random Variable Y** we can map to:

- $\mathbf{Y \rightarrow \{0, 1, 2, 3\}}$

To generalize we write:

- $\mathbf{(Y = y) : y \in \{0, 1, 2, 3\}}$

We assign a Probability as:

- $\mathbf{P(Y = y)}$
- $\mathbf{i.e.\ P(Y = 0), P(Y = 1), P(Y = 2) P(Y = 3)}$

Observation:

In this case a **Single Random Variable** $\mathbf{(Y = y)}$ models three different uncertainty values.

Q: How do you interpret this?

One Interpretation:
A **function** that can take input **0, 1,2,3** and provides the probability measure as an **Output**.

5.3.2 Random Variable in Practice.

Let Y be the number of heads on three coin flips:

0: Head Possible, then Y = 0: Probability of Y on 0 head is given by:
P(Y=0) = 1/8 {Events: (T,T,T)}

1: Head Possible, then Y = 1: Probability of Y on 1 head is given by:
P(Y=1) = 3/8 {Events: (H,T,T), (T,H,T),(T,T,H)}

2: Head Possible, then Y = 2: Probability of Y on 2 head is given by:
P(Y=2) = 3/8 {Events: (H,H,T), (H,T,H),(T,H,H)}

3: Head Possible, then Y = 3: Probability of Y on 3 head is given by:
P(Y=3) = 1/8 {Events: (H,H,H)}

4: Head Possible, then Y = 4: Probability of Y on 4 head is given by:
P(Y≥4) = 0 {Events: ()}

Thus for **Our Random Variable Y** we can map to:

- $\mathbf{Y \rightarrow \{0, 1, 2, 3\}}$

To generalize we write:

- $\mathbf{(Y = y) : y \in \{0, 1, 2, 3\}}$

We assign a Probability as:

- $\mathbf{P(Y = y)}$
- $\mathbf{i.e.\ P(Y = 0), P(Y = 1), P(Y = 2) P(Y = 3)}$

Observation:

In this case a **Single Random Variable** $\mathbf{(Y = y)}$ models three different uncertainty values.

Q: How do you interpret this?

One Interpretation:
A **function** that can take input **0, 1,2,3** and provides the probability measure as an **Output**.

Let's name the function – **Probability Distribution Function**.

5.4 Probability Distribution Function.

- The probability of a random variable (**Discrete or Continuous**) is a **list of probabilities** associated with each of **its possible values**.
  - Probability Distribution for Discrete Random Variables are also known as **Probability Mass Function.**
  - Probability Distribution for Continuous Random Variables are also known as **Probability Density Function**.

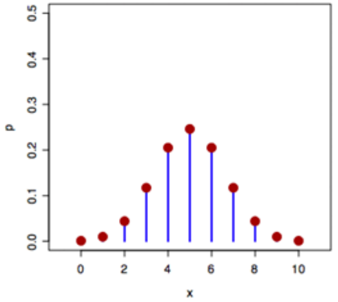


Fig: Discrete Probability Distribution

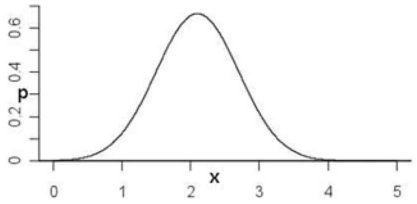


Fig: Continuous Probability Distribution

image from internet : may subjected to copyright.

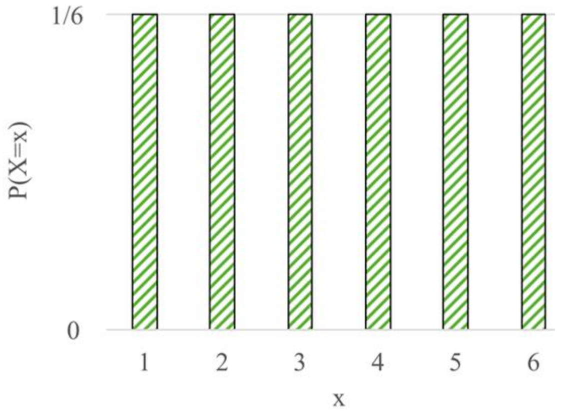
5.5 Probability Mass Function.

- The Probability mass function is probability distribution for a discrete variable Y is a
  - **representation** of all **the possible probabilities** a **random variable Y** can take in the form of :
    - **a formula**, **a table**, or **a graph.**
- The probability distribution for a discrete random variable
  - assigns nonzero probabilities to only a **countable** number of distinct y values.
  - Any value y not explicitly assigned a positive probability is understood to be such that **p(y) = 0**.
- For any discrete probability distribution, the following must be true:
  - **$0 \le p(y) \le 1$ for all y.**
  - $\sum_y \boldsymbol{p(y)} = \mathbf{1}$
    - {where the summation is over all values of y with nonzero probability.}

- Let's illustrate these ideas with an example.
  - PMF for **Y: the value of a six sided die roll**.
  - By the definition:
    - **P(Y = y)** can take values of
      - **P(Y =1); P(Y =2); P(Y =3); P(Y =4); P(Y =5); P(Y =6)**
  - As per definition of Probability Distribution, we can arrange this in table, graph or formula.

Table

Value(y)	1	2	3	4	5	6
P(Y=y)	1/6	1/6	1/6	1/6	1/6	1/6



5.5.1 PMF – Working Example:

- A supervisor in a manufacturing plant has **three men and three women working for him**. He wants to choose **two workers for a special job**. Not wishing to show any biases in his selection, he decides to select the **two workers at random**.
  - Let **Y** denote the **number of women in his selection**. Find the probability distribution for Y .
- *Solution:*

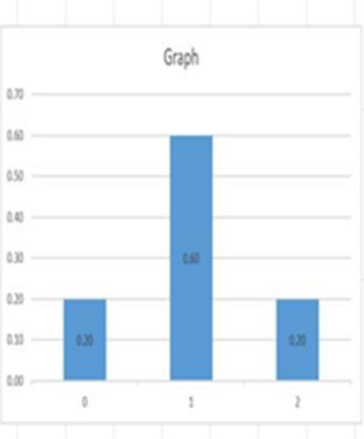
5.5.1 PMF – Working Example – Solution:

- Total possible outcomes ➔
  - Total workers supervisor can select = $\binom{6}{2}$ = 15.
- If **Y = number of women in selection** ,
  - **Case1: zero women were selected:**
    - $\mathbf{P(Y = 0) = \frac{3C0,3C2}{15} = \frac{1}{5}.}$
  - **Case2: One Woman and One Man are Selected:**
    - $\mathrm{P(Y = 1)} = \frac{\mathbf{3C1,3C1}}{\mathbf{15}} = \frac{\mathbf{3}}{\mathbf{5}}.$
  - **Case3: Both the Worker are Women:**
    - $\mathbf{P(Y = 2) = \frac{3C2,3C0}{15} = \frac{1}{5}}.$
- To Find the **Probability Distribution** we have to represent afore probability either in **Equation** or **Table** or **Graph**.
- Our Random Variable Y takes the form:
  - $\mathbf{Y = y: y \in \{0, 1, 2\}}$

- PMF is the representation:
- A equation:
- $\mathbf{P(Y{=}y) = \frac{3Cy,3C_{2-y}}{6C2}}$**; for y = 0, 1,2**
- A table:

y	0	1	2	Σ
P(Y = y)	**0.20**	**0.60**	**0.20**	**1**

- A Graph:

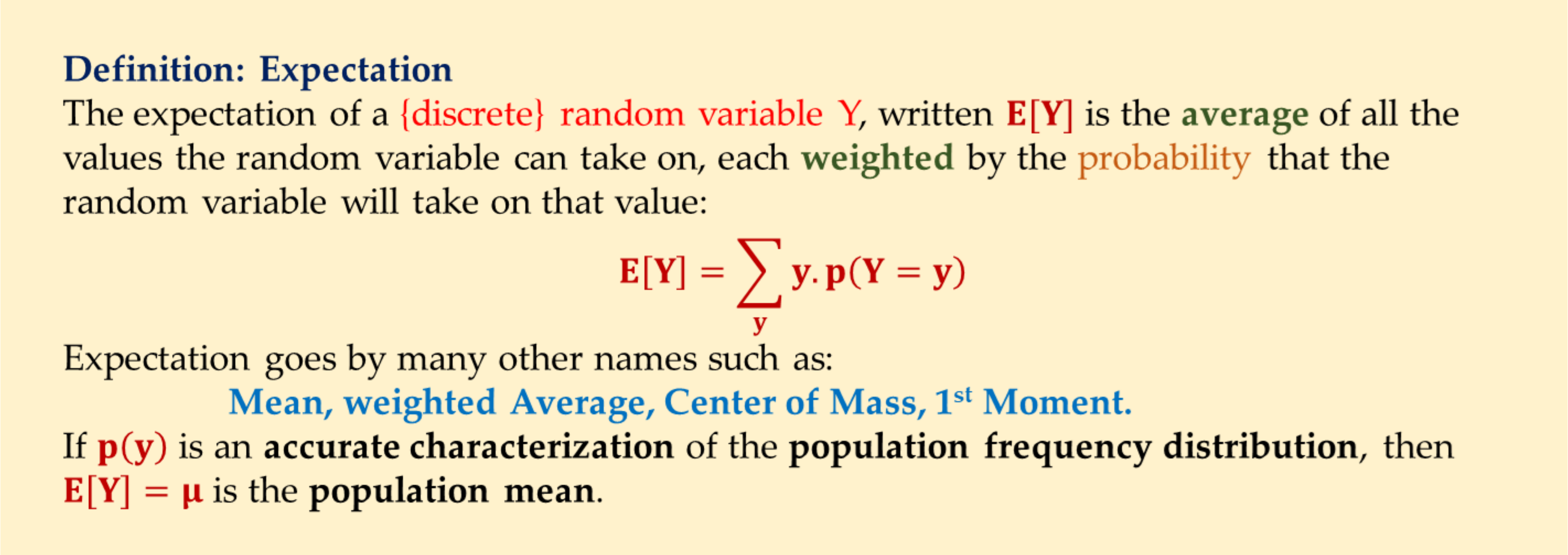


Statistics of a Random Variable.

{6. Mean and Variance of a Random Variable.}

6.1Expectation of Discrete Random Variable.

- A random variable is fully represented by its probability mass function (PMF), which represents each of the values the random variable can take on, and the corresponding probabilities.
- Expectation can be thought as a {**Summary statistics} - - Weighted Average** of a Random Variable!



Definition: Expectation

The expectation of a {discrete} random variable Y, written **E[Y]** is the **average** of all the values the random variable can take on, each **weighted** by the probability that the random variable will take on that value:

$$\mathbf{E[Y]} = \sum_{\mathbf{y}} \mathbf{y.\,p(Y = y)}$$

Expectation goes by many other names such as:

Mean, weighted Average, Center of Mass, 1st Moment.

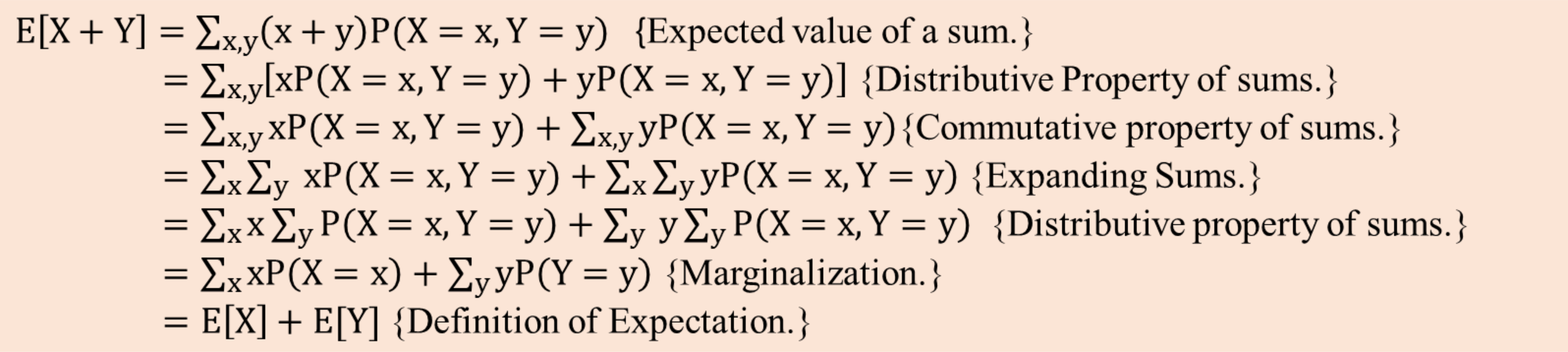
If **p(y)** is an **accurate characterization** of the **population frequency distribution**, then **E[Y] = μ** is the **population mean**.

2.3.1 Properties of Expectations.

- **Property: Linearity of Expectation**
  - $\mathbf{E[aY + b] = aE[Y] + b}$
    - Where a and b are constants and not random variables.
- **Property: Expectation of Constant**
  - $\mathbf{E[a] = a}$
  - Sometimes in proofs, you will end up with the expectation of a constant (rather than a random variable). For example what does the E[5] mean? Since 5 is not a random variable, it does not change, and will always be 5, E[5]=5
- **Property: Law of Unconscious Statistician**
  - $\mathbf{E[g(X)] = \sum_x g(x)P(X = x)}$
  - One can also calculate the expected value of a function g(X) of a random variable X when one knows the probability distribution of X but one does not explicitly know the distribution of g(X). This theorem has the humorous name of "the Law of the Unconscious Statistician" (LOTUS), because it is so useful that you should be able to employ it unconsciously.

2.3.1.1 Properties of Expectations.

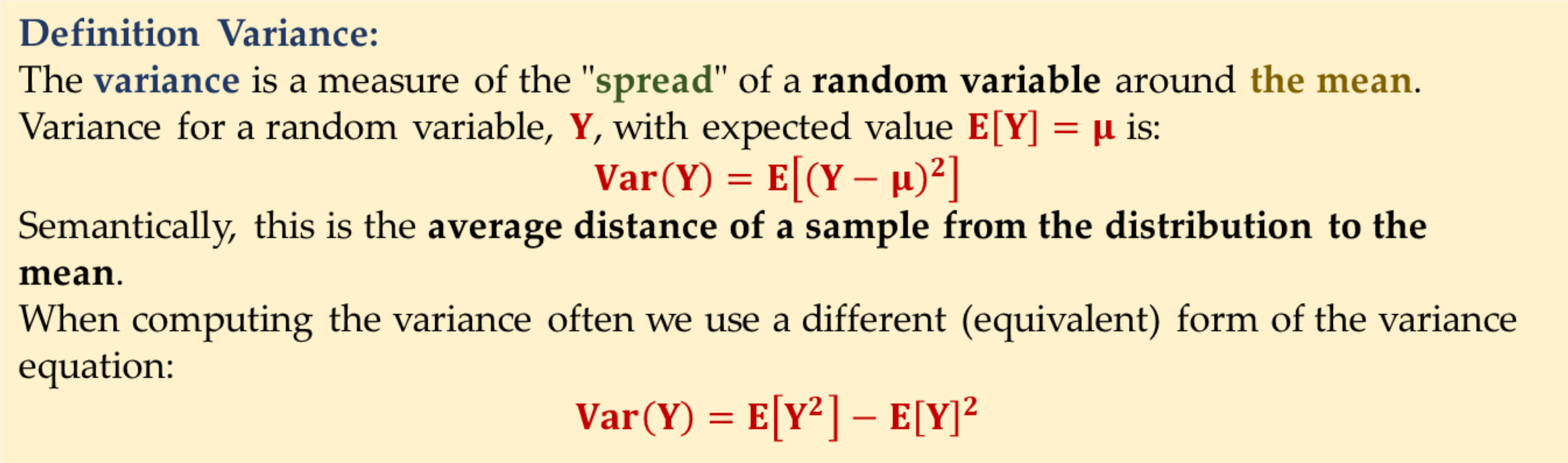
- **Property: Expectation of the Sum of Random Variables**
  - $\mathbf{E[X + Y] = E[X] + E[Y]}$
  - This is true regardless of the relationship between X and Y. They can be dependent, and they can have different distributions. This also applies with more than two random variables.
  - $\mathbf{E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i]}$
- **Proof (Optional – Sample):**



$$\begin{aligned}
E[X+Y] &= \textstyle\sum_{x,y}(x+y)P(X=x,Y=y) \quad \{\text{Expected value of a sum.}\}\\
&= \textstyle\sum_{x,y}[xP(X=x,Y=y)+yP(X=x,Y=y)] \ \{\text{Distributive Property of sums.}\}\\
&= \textstyle\sum_{x,y} xP(X=x,Y=y)+\sum_{x,y} yP(X=x,Y=y)\{\text{Commutative property of sums.}\}\\
&= \textstyle\sum_x\sum_y \ xP(X=x,Y=y)+\sum_x\sum_y yP(X=x,Y=y) \ \{\text{Expanding Sums.}\}\\
&= \textstyle\sum_x x\sum_y P(X=x,Y=y)+\sum_y \ y\sum_y P(X=x,Y=y) \quad \{\text{Distributive property of sums.}\}\\
&= \textstyle\sum_x xP(X=x)+\sum_y yP(Y=y) \ \{\text{Marginalization.}\}\\
&= E[X]+E[Y] \ \{\text{Definition of Expectation.}\}
\end{aligned}$$

2.4 Variance of Discrete Random Variable.

- In the last slide we showed that Expectation was a useful summary of a random variable
  - it calculates the "**weighted average**" of the **random variable**.
- One of the next most important properties of random variables to understand is variance:
  - **the measure of spread**



Definition Variance:
The **variance** is a measure of the "**spread**" of a **random variable** around **the mean**.
Variance for a random variable, $\mathbf{Y}$, with expected value $\mathbf{E[Y] = \mu}$ is:

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

Semantically, this is the **average distance of a sample from the distribution to the mean**.
When computing the variance often we use a different (equivalent) form of the variance equation:

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

7. Distribution From Observed Data.

{Empirical Discrete Distribution}

7.1 Justifying the Need for Distribution Function.

- **Motivating Example – 1:**
  - A lab has 6 computers.
    - Let X denote the number of these computers that are in use during lunch hour –
    - $\mathbf{X = \{x_i | i = 0, 1, 2, 3, 4, 5, 6\}}$
      - And the probability distribution of X is provided in the form of a equation as:
        - $\mathbf{P(X = x) = \frac{x+1}{28}; for\ x = 0, 1, 2, \ldots, 6.}$
          - Here $\mathbf{28 \Rightarrow}$ **Parameter**
  - From here, find:
    - Probability that at most 2 computers are in use.
    - Probability that at least half of the computers are in use.
    - Probability that are 3 or 4 computers free.

7.1.2 Probability Distribution and Parameter.

- Knowing the **distribution function** also ensured we could scale the problem to larger or more complex cases.
- For example, if the total outcomes change (e.g., different computer lab sizes), the same function can still be used to compute probabilities by adjusting its parameters.
    - For instance, if the number of computers changes to 8, the sample space becomes S={0,1,2,…,8}, and the total weight would be:
        - $\sum_{x=0}^{8}(x+1) = 45$
    - The updated distribution would then be
        - $\mathbf{P(X = x) = \frac{x+1}{45}; for\ x = 0, 1, 2, \dots, 8}$
- **Why Parameters are Useful? {General Context}**
    - **Adaptability to Larger Problems**:
        - If the number of computers changes, the parameter 28 (total outcomes) would change accordingly.
        - For instance, if there are 10 computers, the total number of outcomes would become 10+1=11, requiring adjustment in the denominator.
    - **Control Over the Distribution**:
        - By modifying parameters, the distribution can be tailored to reflect different real-world processes.
        - For example, changing the numerator structure to 2x+1 would weight probabilities differently, favoring larger values of x.
    - **Scalability**:
        - Parameters allow the same mathematical model to be used across multiple scenarios, ensuring generalizability without the need to completely redefine the distribution.

7.1.1 Justifying the Need for Distribution Function.

Probability that at most 2 computers are in use $\mathbf{P(X \leq 2)}$**):**
This requires summing the probabilities for **X = 0, 1, and 2**:

- $\mathbf{P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)}$

From the probability formula:

- $\mathbf{P(X = 0) = \frac{0+1}{28} = \frac{1}{28}; P(X = 1) = \frac{1+1}{28} = \frac{2}{28}; P(X = 2) = \frac{2+1}{28} = \frac{3}{28}}$.

Summing:

- $\mathbf{P(X \leq 2) = \frac{1}{28} + \frac{2}{28} + \frac{3}{28} = \frac{3}{14}}$.

- **Our Observation:**
- The **distribution function** gave us an explicit formula to compute **P(X=x) for any x**.
  - Without this, we would need to manually observe data or rely on general assumptions, making the calculations imprecise or impossible.
- Looks like value of probability depends on the parameter 28, which defines our probability distribution function

$$\mathbf{P(X = x) = \frac{x + 1}{28}; for\ x = 0, 1, 2, \ldots, 6.}$$

  - Which in turns depends on a random variable and sample space which in turns depends on our observation which depends on chance experiment.
- If we can identify the parameter of any given distribution function then we may be able to assign probability to all the events mapped by specific random variable.

7.2 Empirical Discrete Distribution: Introduction.

- An *Empirical Discrete Distribution* is a probability distribution derived from observed (empirical) data rather than from a theoretical model.
  - It involves **discrete random variables**, where the possible outcomes are finite or countable.
  - The probabilities are assigned based on the frequency of occurrence of each outcome in a given dataset.
- Here's how it works:
  - **Data Collection**: You start by collecting a set of observed data points. These observations represent the possible outcomes.
  - **Frequency Counting**: For each unique outcome in the data, you count how often it occurs.
  - **Probability Calculation**: The probability of each outcome is calculated by dividing the frequency of the outcome by the total number of observations.

7.3 Empirical Discrete Distribution: Example.

- Suppose you conduct an experiment and observe the following outcomes for the number of times a die shows each face:

Outcome (Face)	Frequency	Probability (P)
1	3	3/30 = 0.10
2	7	7/30 = 0.23
3	5	5/30 = 0.17
4	6	6/30 = 0.20
5	4	4/30 = 0.13
6	5	5/30 = 0.17

  - Here, the empirical discrete distribution assigns probabilities based on observed frequencies, which may differ from the theoretical distribution (**where each face would have a probability of 1/6 or approximately 0.167 for a fair die**).
  - This kind of distribution is useful for **modeling real-world phenomena** where the theoretical distribution is unknown or does not accurately capture the behavior of the observed data.

7.3.1 Empirical Discrete Distribution: Mathematical Formulation.

- Given a set of observed data points $\{\delta_k\}$, where δ_k represents distinct outcomes and $k = 1{:}K$, the empirical discrete distribution is defined using a probability mass function $p_{emp}(A)$where:
  - **Dirac Function**: $\delta(x)$ is a Dirac delta function, which is a function that equals 1 if x is located at a specific point and 0 otherwise.
    - This is used to define whether an observed outcome belongs to a particular subset A.
  - **Weights** w_k: These are non-negative weights associated with each observed outcome δ_k and they must sum up to 1. In other words,
    - $\sum_{k=1}^{K} w_k = 1$
  - **Probability Mass Function:**
    - $p_{emp}(A) = \sum_{k=1}^{K} w_k \delta(A - \delta_k),$
    - where A represents a set of outcomes.
    - The Dirac function:
      - $\delta(A - \delta_k) = 1 \text{ if } \delta_k \in A$
      - $\delta(A - \delta_k) = 0 \text{ if } \delta_k \notin A$

8. Some Special Distributions.

{Discrete Model}

8.1 What and Why of Special Distributions?

- In simple cases (like previous example), **custom probability functions** can suffice, but as the complexity of events increases,
  - studying special distributions **simplifies** the **modeling and computation** of probabilities.
- There are **some specific distributions** that are used over and over in practice, thus they have been given special names.
  - Behind such distribution are a **Random Experiments** which
    - models a lot of real life phenomenon and
    - are frequently used in different applications including Machine Learning,
    - thus we devote a section to study them.
  - Thus, probability distributions act as fundamental tools for quantifying uncertainty in both theoretical and applied contexts.
  - For the **discussed distribution** (Discrete Model):
    - We will try to understand the random experiment,
    - And provide a PMFs and there statistics.

8.2 Random Experiment: Bernoulli Trial.

- **Bernoulli Trial:**
  - Bernoulli trial is an experiment with only two possible outcomes.
  - The two possible outcomes are labelled:
    - Success(S) and Failure (F).
  - The **probability of success** is
    - **P(S) = p**
    - and the **probability of failure** is **P(f) = q = 1-p**
- **Examples:**
  - Tossing a Coin
    - **{Success = H, Failure = T and p = P(S) = P(H)}**
  - Inspecting an item for defects
    - **{Success = defective, Failure = non-defective, and p = P(S) = P(defective)}**

8.3 Bernoulli Distribution: Experiment.

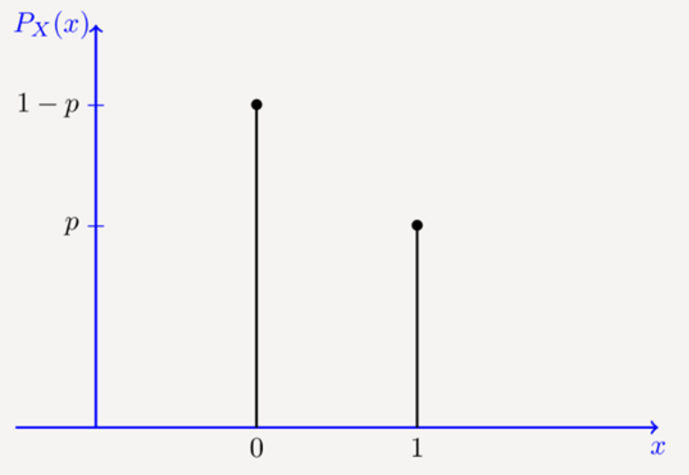
- The Bernoulli (Experiment) Process:
  - Bernoulli Process is an Random Experiment that must satisfy the following properties:
    - The Experiment consists of **n repeated Bernoulli trials**.
    - The **probability of success,{P(S) = p}** remains **constant** from **trial to trial.**
    - The **repeated trials** are **independent**, i.e. the **outcome** on one trial **has no effect** on the **outcome** of any other trial.

8.4.1 Bernoulli Distribution: PMF.

Definition Bernoulli Distribution and its PMF:

A Random Variable **Y** is said to be a Bernoulli random variable with **parameter p**, shown (written) as **Y~ Bernoulli(p)**, if it's PMF is given by:

$$P(Y = y) = \begin{cases} p & \text{for } x = 1 \\ 1 - p & \text{for } x = 0 \\ 0 & \text{Otherwise} \end{cases}$$



← **PMF of a Bernoulli(p) random variable.**

8.4.2 Statistics of Bernoulli.

Expectation of Bernoulli.

- IF Y is a Bernoulli with parameter **p** , i.e.
  - **Y ~ Bernoulli(p),**
- then **Expectation** of Bernoulli Random Variable Y is :
  - $\mathbf{E[Y] = p}$
- Proof – Optional:

$$
\begin{aligned}
\mathbf{E[Y]} &= \mathbf{\textstyle\sum_y y.P(Y = y)} && \textbf{\{Definition of Expectation\}}\\
&= \mathbf{1.p + 0.(1 - p)} && \textbf{\{Y can take on value 0 and 1\}}\\
&= \mathbf{p} && \textbf{\{Remove the zero term\}}
\end{aligned}
$$

Variance of Bernoulli.

- If Y is a Bernoulli with parameter **p**, i.e.
  - **Y ~ Bernoulli(p),**
- with Expectation $\mathbf{E[Y] = p}$
- then **variance** of Bernoulli Random Variable Y is :
  - $\mathbf{Var(Y) = p(1 - p).}$
- Proof – Optional:

First Compute $\mathbf{E[Y^2]}$**:**

$$
\begin{aligned}
\mathbf{E\left[Y^2\right]} &= \mathbf{\textstyle\sum_y y^2.P(Y = y)} && \textbf{\{LOTUS\}}\\
&= \mathbf{0^2.(1 - p) + 1^2.p}\\
&= \mathbf{p}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{Var(Y)} &= \mathbf{E\left[Y^2\right] - E[Y]^2} && \textbf{\{Def. of Variance\}}\\
&= \mathbf{p - p^2} && \mathbf{\{Substitute\ E\left[X^2\right] = p\ and\ E[X] = p.}\\
&= \mathbf{p(1 - p)}
\end{aligned}
$$

8.5 Binomial Distribution: Experiment.

- Consider **n independent (Bernoulli) trials** of an experiment where each trial is a "**success**" with **probability p**.
- Example:
  - For 5 tosses of coin, what is the probability you flip exactly 3 heads. Let X be the random variable, Where X ={0, 1, 2, 3, 4, 5}. Find a PMF.

8.5.1 Binomial Distribution: Experiment.

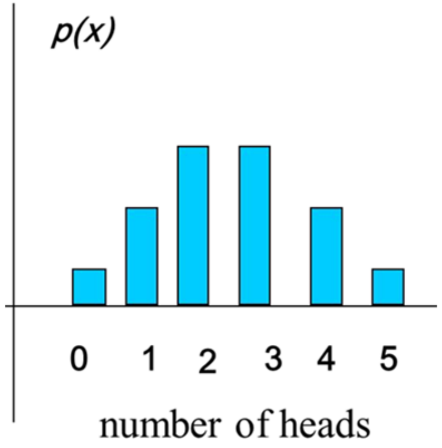
- Consider **n independent (Bernoulli) trials** of an experiment where each trial is a "**success**" with **probability p**.
- Example:
  - For 5 tosses of coin, what is the probability you flip exactly 3 heads. Let X be the random variable, Where X ={0, 1, 2, 3, 4, 5}. Find a PMF.

$\binom{5}{3}$ ways to arrange 3 heads in 5 trials

${}_5C_3 = 5!/3!2! \ = 10$

Outcome	Probability
THHHT	$(1/2)^3 \, x \, (1/2)^2$
HHHTT	$(1/2)^3 \, x \, (1/2)^2$
TTHHH	$(1/2)^3 \, x \, (1/2)^2$
HTTHH	$(1/2)^3 \, x \, (1/2)^2$
HHTTH	$(1/2)^3 \, x \, (1/2)^2$
THTHH	$(1/2)^3 \, x \, (1/2)^2$
HTHTH	$(1/2)^3 \, x \, (1/2)^2$
HHTHT	$(1/2)^3 \, x \, (1/2)^2$
THHTH	$(1/2)^3 \, x \, (1/2)^2$
HTHHT	$(1/2)^3 \, x \, (1/2)^2$
10 arrangements	$x \, (1/2)^3 \, x \, (1/2)^2$

The probability of each unique outcome (note: they are all equal)



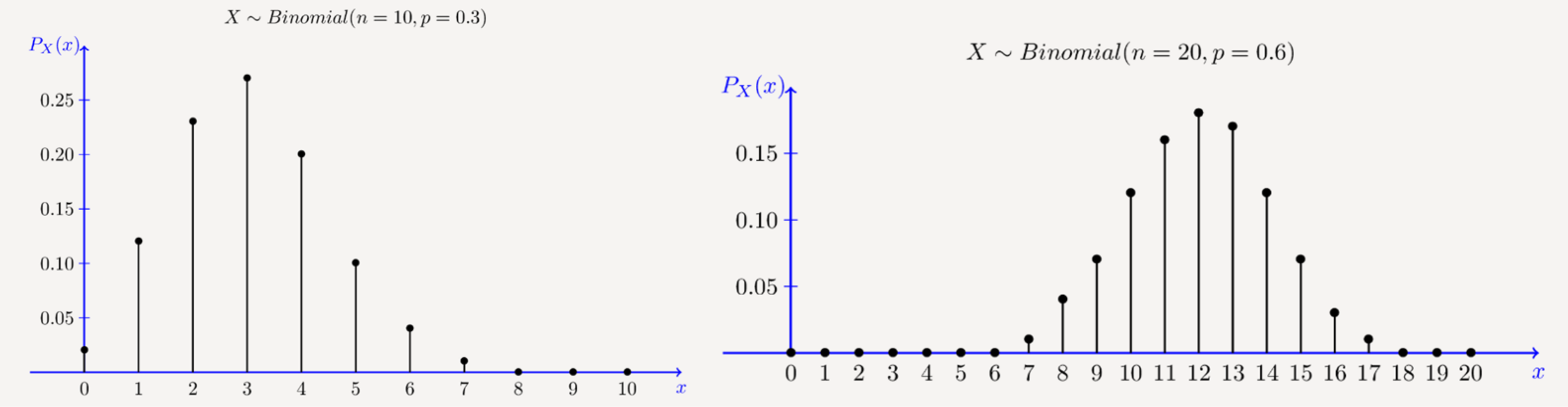
$$P(X=r)=\binom{n}{r}p^r(1-p)^{n-r}$$

8.5.2 Binomial Distribution and PMF.

A random variable Y is said to be a **binomial random variable** with **parameters n and p**, shown as **$Y \sim \text{Binomial}(n, p)$**, if its PMF is given by:

$$P_Y(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } k = 0, 1, 2 \ldots, n \\ 0 & \text{Otherwise} \end{cases}$$

Where $0 < p < 1$.



8.5.3 Bernoulli and Binomial Distribution.

- Relationship is defined by:
  - If $\mathbf{Y_1, Y_2 \dots, Y_n}$ are **independent Bernoulli(p) random variables**, then the random variable X defined by $\mathbf{X = Y_1 + Y_2 + \cdots + Y_n}$ has a **Binomial (n, p) distribution**.
    - $\mathbf{X \sim Bin(n, p) = \sum_{i=1}^{n} Y_i \sim Bern(p)}$

- **Expectation and Variance of Binomial Distribution:**
  - Expectation:
    - $\mathbf{E[Y] = n.p}$
  - Variance:
    - $\mathbf{Var(Y) = n.p.(1 - p)}$

Thank You