EE 178 Lecture Notes 0
Course Introduction

- About EE178
- About Probability
- Course Goals
- Course Topics
- Lecture Notes

EE 178

- EE 178 provides an introduction to probabilistic system analysis:
  - Basic probability theory
  - Some statistics
  - Examples from EE systems
- Fulfills the EE undergraduate probability and statistics requirement (alternatives include Stat 116 and Math 151)
- Very important background for EEs
  - Signal processing
  - Communication
  - Communication networks
  - Reliability of systems
  - Noise in circuits
• Also important in *many* other fields, including
  ○ Computer science (Artificial intelligence)
  ○ Economics
  ○ Finance (stock market)
  ○ Physics (statistical physics, quantum physics)
  ○ Medicine (drug trials)

• Knowledge of probability is almost as necessary as calculus and linear algebra

---

**Probability Theory**

• Probability provides mathematical models for random phenomena and experiments, such as: gambling, stock market, packet transmission in networks, electron emission, noise, statistical mechanics, etc.

• Its origin lies in observations associated with games of chance:
  ○ If an "unbiased" coin is tossed $n$ times and heads come up $n_H$ times, the relative frequency of heads, $n_H/n$, is likely to be very close to $1/2$
  ○ If a card is drawn from a perfectly shuffled deck and then replaced, the deck is reshuffled, and the process is repeated $n$ times, the relative frequency of spades is around $1/4$

• The purpose of probability theory is to describe and predict such relative frequencies (averages) in terms of probabilities of events

• But how is probability defined?
• Classical definition of probability: This was the prevailing definition for many centuries.
Define the probability of an event \( A \) as:

\[
P(A) = \frac{N_A}{N},
\]

where \( N \) is the number of possible outcomes of the random experiment and \( N_A \) is the number of outcomes favorable to the event \( A \)
For example for a 6-sided die there are 6 outcomes and 3 of them are even, thus
\( P(\text{even}) = \frac{3}{6} \)
Problems with this classical definition:

○ Here the assumption is that all outcomes are equally likely (probable). Thus, the concept of probability is used to define probability itself! Cannot be used as basis for a mathematical theory
○ In many random experiments, the outcomes are not equally likely
○ The definition doesn’t work when the number of possible outcomes is infinite

• Relative frequency definition of probability: Here the probability of an event \( A \) is defined as:

\[
P(A) = \lim_{n \to \infty} \frac{n_A}{n},
\]

where \( n_A \) is the number of times \( A \) occurs in \( n \) performances of the experiment
Problems with the frequency definition:

○ How do we assert that such limit exists?
○ We often cannot perform the experiment multiple times (or even once, e.g., defining the probability of a nuclear plant failure)
The axiomatic definition of probability was introduced by Kolmogorov in 1933. It provides rules for assigning probabilities to events in a mathematically consistent way and for deducing probabilities of events from probabilities of other events.

**Elements of axiomatic definition:**
- Set of all possible outcomes of the random experiment $\Omega$ (Sample space)
- Set of events, which are subsets of $\Omega$
- A probability law (measure or function) that assigns probabilities to events such that:
  1. $P(A) \geq 0$
  2. $P(\Omega) = 1$, and
  3. If $A$ and $B$ are disjoint events, i.e., $A \cap B = \emptyset$, then

\[
P(A \cup B) = P(A) + P(B)
\]

These rules are consistent with relative frequency interpretation.

Unlike the relative frequency definition, where the limit is assumed to exist, the axiomatic approach itself is used to prove when and under what conditions such limit exists (Laws of Large Numbers) and how fast it is approached (Central Limit Theorem).

The theory provides mathematically precise ways for dealing with experiments with infinite number of possible outcomes, defining random variables, etc.

The theory **does not** deal with what the values of the probabilities are or how they are obtained. Any assignment of probabilities that satisfies the axioms is legitimate.

How do we find the actual probabilities? They are obtained from:
- Physics, e.g., Boltzmann, Bose-Einstein, Fermi-Dirac statistics
- Engineering models, e.g., queuing models for communication networks, models for noisy communication channels
- Social science theories, e.g., intelligence, behavior
- Empirically. Collect data and fit a model (may have no physical, engineering, or social meaning ...)

Statistics and Statistical Signal Processing

• Statistics applies probability theory to real world problems. It deals with:
  ○ Methods for data collection and construction of experiments described by
    probabilistic models, and
  ○ Methods for making inferences from data, e.g., estimating probabilities,
    deciding if a drug is effective from drug trials

• Statistical signal processing:
  ○ Develops probabilistic models for EE signals, e.g., speech, audio, images,
    video, and systems, e.g., communication, storage, compression, error
    correction, pattern recognition
  ○ Uses tools from probability and statistics to design signal processing
    algorithms to extract, predict, estimate, detect, compress, . . . such signals
  ○ The field has provided the foundation for the Information Age. Its impact on
    our lives has been huge: cell phones, DSL, digital cameras, DVD, MP3
    players, digital TV, medical imaging, etc.

Course Goals

• Provide an introduction to probability theory (precise but not overly
  mathematical)

• Provide, through examples, some applications of probability to EE systems

• Provide some exposure to statistics and statistical signal processing

• Provide background for more advanced courses in statistical signal processing
  and communication (e.g., EE 278, 279)

• Have a lot of fun solving cool probability problems

  WARNING: To really understand probability you have to work homework
  problems yourself. Just reading solutions or having others give you solutions
  will not do
1. Sample Space and probability (B&T: Chapter 1)
   • Set theory
   • Sample space: discrete, continuous
   • Events
   • Probability law
   • Conditional probability; chain rule
   • Law of total probability
   • Bayes rule
   • Independence
   • Counting

2. Random Variables (B&T: 2.1–2.4, 3.1–3.3, 3.6 pp. 179–186)
   • Basic concepts
   • Probability mass function (pmf)
   • Famous discrete random variables
   • Probability density function (pdf)
   • Famous continuous random variables
   • Mean and variance
   • Cumulative distribution function (cdf)
   • Functions of random variables

3. Multiple Random variables (B&T: 2.5–2.7, 3.4–3.5, 3.6 pp. 186–190, 4.2)
   • Joint, marginal, and conditional pmf: Bayes rule for pmfs, independence
   • Joint, marginal, and conditional pdf: Bayes rule for pdfs, independence
   • Functions of multiple random variables
4. Expectation: (B&T: 2.4, 3.1, pp. 144–148, 4.3, 4.5, 4.6)
   - Definition and properties
   - Correlation and covariance
   - Sum of random variables
   - Conditional expectation
   - Iterated expectation
   - Mean square error estimation

5. Transforms (B&T: 4.1)

6. Limit Theorems (B&T: 7.1, 7.2, 4.1, 7.4)
   - Sample mean
   - Markov and Chebychev inequalities
   - Weak law of large numbers
   - The Central Limit Theorem
   - Confidence Intervals

7. Random Processes: (B&T: 5.1, 6.1–6.4)
   - Basic concepts
   - Bernoulli process
   - Markov chains
Lecture Notes

- Based on B&T text, Prof. Gray EE 178 notes, and other sources
- Help organize the material and reduce note taking in lectures
- You will need to take some notes, e.g. clarifications, missing steps in derivations, solutions to additional examples
- Slide title indicates a topic that often continues over several consecutive slides
- Lecture notes + your notes + review sessions should be sufficient (B&T is only recommended)
Set Theory Basics

- A set is a collection of objects, which are its *elements*
  - $\omega \in A$ means that $\omega$ is an element of the set $A$
  - A set with no elements is called the *empty set*, denoted by $\emptyset$
- Types of sets:
  - Finite: $A = \{\omega_1, \omega_2, \ldots, \omega_n\}$
  - Countably infinite: $A = \{\omega_1, \omega_2, \ldots\}$, e.g., the set of integers
  - Uncountable: A set that takes a continuous set of values, e.g., the $[0, 1]$ interval, the real line, etc.
- A set can be described by all $\omega$ having a certain property, e.g., $A = [0, 1]$ can be written as $A = \{\omega : 0 \leq \omega \leq 1\}$
- A set $B \subset A$ means that every element of $B$ is an element of $A$
- A *universal set* $\Omega$ contains all objects of particular interest in a particular context, e.g., sample space for random experiment
Set Operations

• Assume a universal set $\Omega$

• Three basic operations:
  ○ Complementation: A complement of a set $A$ with respect to $\Omega$ is $A^c = \{\omega \in \Omega : \omega \notin A\}$, so $\Omega^c = \emptyset$
  ○ Intersection: $A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}$
  ○ Union: $A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}$

• Notation:
  ○ $\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \ldots \cup A_n$
  ○ $\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \ldots \cap A_n$

• A collection of sets $A_1, A_2, \ldots, A_n$ are disjoint or mutually exclusive if $A_i \cap A_j = \emptyset$ for all $i \neq j$, i.e., no two of them have a common element

• A collection of sets $A_1, A_2, \ldots, A_n$ partition $\Omega$ if they are disjoint and $\bigcup_{i=1}^{n} A_i = \Omega$

• Venn Diagrams

(a) $\Omega$

(b) $A$

(c) $B$

(d) $B^c$

(e) $A \cap B$

(f) $A \cup B$
Algebra of Sets

- Basic relations:
  1. \( S \cap \Omega = S \)
  2. \((A^c)^c = A\)
  3. \(A \cap A^c = \emptyset\)
  4. Commutative law: \( A \cup B = B \cup A \)
  5. Associative law: \( A \cup (B \cup C) = (A \cup B) \cup C \)
  6. Distributive law: \( A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \)
  7. DeMorgan's law: \((A \cap B)^c = A^c \cup B^c\)
    DeMorgan's law can be generalized to \(n\) events:
    \[
    (\bigcap_{i=1}^n A_i)^c = \bigcup_{i=1}^n A_i^c
    \]

- These can all be proven using the definition of set operations or visualized using Venn Diagrams

Elements of Probability

- Probability theory provides the mathematical rules for assigning probabilities to outcomes of random experiments, e.g., coin flips, packet arrivals, noise voltage

- Basic elements of probability:
  - **Sample space**: The set of all possible “elementary” or “finest grain” outcomes of the random experiment (also called *sample points*)
    - The sample points are all *disjoint*
    - The sample points are *collectively exhaustive*, i.e., together they make up the entire sample space
  - **Events**: Subsets of the sample space
  - **Probability law**: An assignment of probabilities to events in a mathematically consistent way
Discrete Sample Spaces

- Sample space is called *discrete* if it contains a countable number of sample points

- Examples:
  - Flip a coin once: $\Omega = \{H, T\}$
  - Flip a coin three times: $\Omega = \{HHH, HHT, HTH, \ldots\} = \{H, T\}^3$
  - Flip a coin $n$ times: $\Omega = \{H, T\}^n$ (set of sequences of H and T of length $n$)
  - Roll a die once: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Roll a die twice: $\Omega = \{(1, 1), (1, 2), (2, 1), \ldots, (6, 6)\} = \{1, 2, 3, 4, 5, 6\}^2$
  - Flip a coin until the first heads appears: $\Omega = \{H, TH, TTH, TTTH, \ldots\}$
  - Number of packets arriving in time interval $(0, T]$ at a node in a communication network: $\Omega = \{0, 1, 2, 3, \ldots\}$

  Note that the first five examples have *finite* $\Omega$, whereas the last two have *countably infinite* $\Omega$. Both types are called discrete

- Sequential models: For sequential experiments, the sample space can be described in terms of a tree, where each outcome corresponds to a terminal node (or a *leaf*)

  Example: Three flips of a coin
- Example: Roll a fair four-sided die twice.
  Sample space can be represented by a tree as above, or graphically

```
1 2 3 4
```

- **Continuous Sample Spaces**
  - A *continuous* sample space consists of a continuum of points and thus contains an uncountable number of points
  
  - Examples:
    - Random number between 0 and 1: \( \Omega = [0, 1] \)
    - Suppose we pick two numbers at random between 0 and 1, then the sample space consists of all points in the *unit square*, i.e., \( \Omega = [0, 1]^2 \)
Packet arrival time: $t \in (0, \infty)$, thus $\Omega = (0, \infty)$

Arrival times for $n$ packets: $t_i \in (0, \infty)$, for $i = 1, 2, \ldots, n$, thus $\Omega = (0, \infty)^n$

- A sample space is said to be *mixed* if it is neither discrete nor continuous, e.g., $\Omega = [0, 1] \cup \{3\}$

Events

- Events are *subsets* of the sample space. An event occurs if the outcome of the experiment belongs to the event

Examples:
- Any outcome (sample point) is an event (also called an elementary event), e.g., $\{HTH\}$ in three coin flips experiment or $\{0.35\}$ in the picking of a random number between 0 and 1 experiment
- Flip coin 3 times and get exactly one H. This is a more complicated event, consisting of three sample points $\{TTH, THT, HTT\}$
- Flip coin 3 times and get an odd number of H’s. The event is $\{TTH, THT, HTT, HHH\}$
- Pick a random number between 0 and 1 and get a number between 0.0 and 0.5. The event is $[0, 0.5]$.
- An event might have *no points* in it, i.e., be the empty set $\emptyset$
Axioms of Probability

- Probability law (measure or function) is an assignment of probabilities to events (subsets of sample space \( \Omega \)) such that the following three axioms are satisfied:

1. \( P(A) \geq 0 \), for all \( A \) (nonnegativity)
2. \( P(\Omega) = 1 \) (normalization)
3. If \( A \) and \( B \) are disjoint (\( A \cap B = \emptyset \)), then

\[
P(A \cup B) = P(A) + P(B) \quad \text{(additivity)}
\]

More generally,

3’. If the sample space has an infinite number of points and \( A_1, A_2, \ldots \) are disjoint events, then

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

- Mimics relative frequency, i.e., perform the experiment \( n \) times (e.g., roll a die \( n \) times). If the number of occurrences of \( A \) is \( n_A \), define the relative frequency of an event \( A \) as \( f_A = n_A/n \)
  - Probabilities are nonnegative (like relative frequencies)
  - Probability something happens is 1 (again like relative frequencies)
  - Probabilities of disjoint events add (again like relative frequencies)
- Analogy: Except for normalization, probability is a measure much like
  - mass
  - length
  - area
  - volume

They all satisfy axioms 1 and 3

This analogy provides some intuition but is not sufficient to fully understand probability theory — other aspects such as conditioning, independence, etc., are unique to probability
Probability for Discrete Sample Spaces

- Recall that sample space \( \Omega \) is said to be *discrete* if it is countable.

- The probability measure \( P \) can be simply defined by first assigning probabilities to outcomes, i.e., elementary events \( \{ \omega \} \), such that:
  \[
P(\{\omega\}) \geq 0, \text{ for all } \omega \in \Omega, \text{ and } \sum_{\omega \in \Omega} P(\{\omega\}) = 1
  \]

- The probability of any other event \( A \) (by the additivity axiom) is simply
  \[
P(A) = \sum_{\omega \in A} P(\{\omega\})
  \]

- Examples:
  - For the coin flipping experiment, assign \( P(\{\text{H}\}) = p \) and \( P(\{\text{T}\}) = 1 - p \), for \( 0 \leq p \leq 1 \)
    Note: \( p \) is the *bias* of the coin, and a coin is *fair* if \( p = \frac{1}{2} \)
  - For the die rolling experiment, assign \( P(\{i\}) = \frac{1}{6} \), for \( i = 1, 2, \ldots, 6 \)
    The probability of the event "the outcome is even", \( A = \{2, 4, 6\} \), is
    \[
P(A) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{2}
    \]
○ If $\Omega$ is countably infinite, we can again assign probabilities to elementary events.

Example: Assume $\Omega = \{1, 2, \ldots\}$, assign probability $2^{-k}$ to event $\{k\}$.

The probability of the event "the outcome is even"

$$P(\text{outcome is even}) = P(\{2, 4, 6, 8, \ldots\})$$

$$= P(\{2\}) + P(\{4\}) + P(\{6\}) + \ldots$$

$$= \sum_{k=1}^{\infty} P(\{2k\})$$

$$= \sum_{k=1}^{\infty} 2^{-2k} = \frac{1}{3}$$

## Probability for Continuous Sample Space

- Recall that if a sample space is continuous, $\Omega$ is uncountably infinite.
- For continuous $\Omega$, we cannot in general define the probability measure $P$ by first assigning probabilities to outcomes.
- To see why, consider assigning a uniform probability measure to $\Omega = (0, 1]$.
  - In this case the probability of each single outcome event is zero.
  - How do we find the probability of an event such as $A = \left[\frac{1}{2}, \frac{3}{4}\right]$?
- For this example we can define uniform probability measure over $[0, 1]$ by assigning to an event $A$, the probability

$$P(A) = \text{length of } A,$$

e.g., $P([0, 1/3] \cup [2/3, 1]) = 2/3$

Check that this is a legitimate assignment.
Another example: Romeo and Juliet have a date. Each arrives late with a random delay of up to 1 hour. Each will wait only 1/4 of an hour before leaving. What is the probability that Romeo and Juliet will meet?

Solution: The pair of delays is equivalent to that achievable by picking two random numbers between 0 and 1. Define probability of an event as its area. The event of interest is represented by the cross hatched region:

Probability of the event is:

\[
\text{area of crosshatched region} = 1 - 2 \times \frac{1}{2}(0.75)^2 = 0.4375
\]

Basic Properties of Probability

- There are several useful properties that can be derived from the axioms of probability:
  1. \( P(A^c) = 1 - P(A) \)
     - \( P(\emptyset) = 0 \)
     - \( P(A) \leq 1 \)
  2. If \( A \subseteq B \), then \( P(A) \leq P(B) \)
  3. \( P(A \cup B) = P(A) + P(B) - P(A \cap B) \)
  4. \( P(A \cup B) \leq P(A) + P(B) \), or in general

\[
P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)
\]

This is called the \textit{Union of Events Bound}.

- These properties can be proved using the axioms of probability and visualized using Venn diagrams.
Conditional Probability

- Conditional probability allows us to compute probabilities of events based on partial knowledge of the outcome of a random experiment.

- Examples:
  - We are told that the sum of the outcomes from rolling a die twice is 9. What is the probability the outcome of the first die was a 6?
  - A spot shows up on a radar screen. What is the probability that there is an aircraft?
  - You receive a 0 at the output of a digital communication system. What is the probability that a 0 was sent?

- As we shall see, conditional probability provides us with two methods for computing probabilities of events: the sequential method and the divide-and-conquer method.

- It is also the basis of inference in statistics: make an observation and reason about the cause.

- In general, given an event \( B \) has occurred, we wish to find the probability of another event \( A \), \( P(A|B) \).

- If all elementary outcomes are equally likely, then

\[
P(A|B) = \frac{\text{# of outcomes in both } A \text{ and } B}{\text{# of outcomes in } B}
\]

- In general, if \( B \) is an event such that \( P(B) \neq 0 \), the conditional probability of any event \( A \) given \( B \) is defined as

\[
P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{or} \quad \frac{P(A, B)}{P(B)}
\]

- The function \( P(\cdot|B) \) for fixed \( B \) specifies a probability law, i.e., it satisfies the axioms of probability.
Example

• Roll a fair four-sided die twice. So, the sample space is \( \{1, 2, 3, 4\}^2 \). All sample points have probability 1/16

Let \( B \) be the event that the minimum of the two die rolls is 2 and \( A_m \), for \( m = 1, 2, 3, 4 \), be the event that the maximum of the two die rolls is \( m \). Find \( P(A_m|B) \)

• Solution:

\begin{align*}
\text{2nd roll} & \quad \text{1st roll} \\
\begin{array}{cccc}
4 & 3 & 2 & 1 \\
3 & 2 & 1 & 1 \\
2 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
\end{array}
\end{align*}

\begin{align*}
P(A_m|B) & \quad m \\
\begin{array}{cccc}
& 4 & 3 & 2 & 1 \\
1 & 1 & 1 & 1 & 1 \\
2 & 1 & 1 & 1 & 1 \\
3 & 1 & 1 & 1 & 1 \\
4 & 1 & 1 & 1 & 1 \\
\end{array}
\end{align*}

Conditional Probability Models

• Before: Probability law \( \Rightarrow \) conditional probabilities

• Reverse is often more natural: Conditional probabilities \( \Rightarrow \) probability law

• We use the \textit{chain rule} (also called \textit{multiplication rule}):

By the definition of conditional probability, \( P(A \cap B) = P(A|B)P(B) \). Suppose that \( A_1, A_2, \ldots, A_n \) are events, then
\[
P(A_1 \cap A_2 \cap A_3 \cdots \cap A_n)
= P(A_1 \cap A_2 \cap A_3 \cdots \cap A_{n-1}) \times P(A_n | A_1 \cap A_2 \cap A_3 \cdots \cap A_{n-1})
= P(A_1 \cap A_2 \cap A_3 \cdots \cap A_{n-2}) \times P(A_n | A_1 \cap A_2 \cap A_3 \cdots \cap A_{n-2})
\times P(A_n | A_1 \cap A_2 \cap A_3 \cdots \cap A_{n-1})
\vdots
= P(A_1) \times P(A_2 | A_1) \times P(A_3 | A_1 \cap A_2) \cdots P(A_n | A_1 \cap A_2 \cap A_3 \cdots \cap A_{n-1})
= \prod_{i=1}^{n} P(A_i | A_1, A_2, \ldots, A_{i-1}),
\]

where \( A_0 = \emptyset \)

---

**Sequential Calculation of Probabilities**

- **Procedure:**
  1. Construct a tree description of the sample space for a sequential experiment
  2. Assign the conditional probabilities on the corresponding branches of the tree
  3. By the chain rule, the probability of an outcome can be obtained by multiplying the conditional probabilities along the path from the root to the leaf node corresponding to the outcome

- **Example (Radar Detection):** Let \( A \) be the event that an aircraft is flying above and \( B \) be the event that the radar detects it. Assume \( P(A) = 0.05 \), \( P(B|A) = 0.99 \), and \( P(B|A^c) = 0.1 \)

  What is the probability of
  - Missed detection?, i.e., \( P(A \cap B^c) \)
  - False alarm?, i.e., \( P(B \cap A^c) \)

  The sample space is: \( \Omega = \{ (A, B), (A^c, B), (A, B^c), (A^c, B^c) \} \)
Solution: Represent the sample space by a tree with conditional probabilities on its edges

Example: Three cards are drawn at random (without replacement) from a deck of cards. What is the probability of not drawing a heart?

Solution: Let $A_i$, $i = 1, 2, 3$, represent the event of no heart in the $i$th draw. We can represent the sample space as:

$$
\Omega = \{(A_1, A_2, A_3), (A_1^c, A_2, A_3), \ldots, (A_1^c, A_2^c, A_3^c)\}
$$

To find the probability law, we represent the sample space by a tree, write conditional probabilities on branches, and use the chain rule.
Total probability – Divide and Conquer Method

- Let $A_1, A_2, \ldots, A_n$ be events that partition $\Omega$, i.e., that are disjoint ($A_i \cap A_j = \emptyset$ for $i \neq j$) and $\bigcup_{i=1}^{n} A_i = \Omega$. Then for any event $B$

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

  This is called the Law of Total Probability. It also holds for $n = \infty$. It allows us to compute the probability of a complicated event from knowledge of probabilities of simpler events.

- Example: Chess tournament, 3 types of opponents for a certain player.
  - P(Type 1) = 0.5, P(Win | Type1) = 0.3
  - P(Type 2) = 0.25, P(Win | Type2) = 0.4
  - P(Type 3) = 0.25, P(Win | Type3) = 0.5

  What is probability of player winning?
Solution: Let $B$ be the event of winning and $A_i$ be the event of playing Type $i$, $i = 1, 2, 3$:

$$P(B) = \sum_{i=1}^{3} P(A_i)P(B|A_i)$$

$$= 0.5 \times 0.3 + 0.25 \times 0.4 + 0.25 \times 0.5$$

$$= 0.375$$

---

**Bayes Rule**

- Let $A_1, A_2, \ldots, A_n$ be nonzero probability events (the causes) that partition $\Omega$, and let $B$ be a nonzero probability event (the effect)

- We often know the *a priori* probabilities $P(A_i)$, $i = 1, 2, \ldots, n$ and the conditional probabilities $P(B|A_i)$s and wish to find the *a posteriori probabilities* $P(A_j|B)$ for $j = 1, 2, \ldots, n$

- From the definition of conditional probability, we know that

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

By the law of total probability

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$
Substituting we obtain *Bayes rule*

\[ P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)} \text{ for } j = 1, 2, \ldots, n \]

- Bayes rule also applies when the number of events \( n = \infty \)

- Radar Example: Recall that \( A \) is event that the aircraft is flying above and \( B \) is the event that the aircraft is detected by the radar. What is the probability that an aircraft is actually there given that the radar indicates a detection?

Recall \( P(A) = 0.05 \), \( P(B|A) = 0.99 \), \( P(B|A^c) = 0.1 \). Using Bayes rule:

\[
P(\text{there is an aircraft|radar detects it}) = P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}
\]

\[
= \frac{0.05 \times 0.99}{0.05 \times 0.99 + 0.95 \times 0.1}
\]

\[
= 0.3426
\]

---

### Binary Communication Channel

- Consider a noisy binary communication channel, where 0 or 1 is sent and 0 or 1 is received. Assume that 0 is sent with probability 0.2 (and 1 is sent with probability 0.8)

  The channel is noisy. If a 0 is sent, a 0 is received with probability 0.9, and if a 1 is sent, a 1 is received with probability 0.975

- We can represent this channel model by a *probability transition diagram*

![Probability Transition Diagram](image)

Given that 0 is received, find the probability that 0 was sent
• This is a random experiment with sample space \( \Omega = \{(0, 0), (0, 1), (1, 0), (1, 1)\} \), where the first entry is the bit sent and the second is the bit received.

• Define the two events

\[
A = \{0 \text{ is sent}\} = \{(0, 1), (0, 0)\}, \quad \text{and} \quad B = \{0 \text{ is received}\} = \{(0, 0), (1, 0)\}
\]

• The probability measure is defined via the \( P(A), P(B|A), \) and \( P(B^c|A^c) \) provided on the probability transition diagram of the channel.

• To find \( P(A|B) \), the \textit{a posteriori} probability that a 0 was sent. We use Bayes rule

\[
P(A|B) = \frac{P(B|A)P(A)}{P(A)P(B|A) + P(A^c)P(B|A^c)},
\]

to obtain

\[
P(A|B) = \frac{0.9}{0.2} \times 0.2 = 0.9,
\]

which is much higher than the \textit{a priori} probability of \( A (= 0.2) \)

---

**Independence**

• It often happens that the knowledge that a certain event \( B \) has occurred has no effect on the probability that another event \( A \) has occurred, i.e.,

\[
P(A|B) = P(A)
\]

In this case we say that the two events are statistically independent.

• Equivalently, two events are said to be \textit{statistically independent} if

\[
P(A, B) = P(A)P(B)
\]

So, in this case, \( P(A|B) = P(A) \) and \( P(B|A) = P(B) \).

• Example: Assuming that the binary channel of the previous example is used to send two bits independently, what is the probability that both bits are in error?
Solution:

- Define the two events
  
  \[ E_1 = \{ \text{First bit is in error} \} \]
  
  \[ E_2 = \{ \text{Second bit is in error} \} \]

- Since the bits are sent independently, the probability that both are in error is
  
  \[ P(E_1, E_2) = P(E_1)P(E_2) \]

  Also by symmetry, \( P(E_1) = P(E_2) \)

  To find \( P(E_1) \), we express \( E_1 \) in terms of the events \( A \) and \( B \) as
  
  \[ E_1 = (A_1 \cap B_1^c) \cup (A_1^c \cap B_1) \]

- Thus,
  
  \[ P(E_1) = P(A_1, B_1^c) + P(A_1^c, B_1) \]
  
  \[ = P(A_1)P(B_1^c | A_1) + P(A_1^c)P(B_1 | A_1^c) \]
  
  \[ = 0.2 \times 0.1 + 0.8 \times 0.025 = 0.04 \]

- The probability that the two bits are in error
  
  \[ P(E_1, E_2) = (0.04)^2 = 16 \times 10^{-4} \]

- In general, \( A_1, A_2, \ldots, A_n \) are mutually independent if for each subset of the events \( A_{i_1}, A_{i_2}, \ldots, A_{i_k} \)
  
  \[ P(A_{i_1}, A_{i_2}, \ldots, A_{i_k}) = \prod_{j=1}^{k} P(A_{i_j}) \]

- Note: \( P(A_1, A_2, \ldots, A_n) = \prod_{j=1}^{n} P(A_i) \) alone is not sufficient for independence
Example: Roll two fair dice independently. Define the events

\[ A = \{ \text{First die } = 1, 2, \text{ or } 3 \} \]
\[ B = \{ \text{First die } = 2, 3, \text{ or } 6 \} \]
\[ C = \{ \text{Sum of outcomes } = 9 \} \]

Are \( A, B, \) and \( C \) independent?

Solution:
Since the dice are fair and the experiments are performed independently, the probability of any pair of outcomes is \( \frac{1}{36} \), and

\[ P(A) = \frac{1}{2}, \quad P(B) = \frac{1}{2}, \quad P(C) = \frac{1}{9} \]

Since \( A \cap B \cap C = \{(3, 6)\} \), \( P(A, B, C) = \frac{1}{36} = P(A)P(B)P(C) \)

But are \( A, B, \) and \( C \) independent? Let’s find

\[ P(A, B) = \frac{2}{6} = \frac{1}{3} \neq \frac{1}{4} = P(A)P(B), \]

and thus \( A, B, \) and \( C \) are not independent!
Also, independence of subsets does not necessarily imply independence

Example: Flip a fair coin twice independently. Define the events:

- \( A \): First toss is Head
- \( B \): Second toss is Head
- \( C \): First and second toss have different outcomes

\( A \) and \( B \) are independent, \( A \) and \( C \) are independent, and \( B \) and \( C \) are independent

Are \( A, B, C \) mutually independent?

Clearly not, since if you know \( A \) and \( B \), you know that \( C \) could not have occurred, i.e., \( P(A, B, C) = 0 \neq P(A)P(B)P(C) = 1/8 \)

---

**Counting**

- Discrete uniform law:
  - Finite sample space where all sample points are equally probable:
    
    \[ P(A) = \frac{\text{number of sample points in } A}{\text{total number of sample points}} \]
  
  - Variation: all outcomes in \( A \) are equally likely, each with probability \( p \). Then,
    
    \[ P(A) = p \times (\text{number of elements of } A) \]

- In both cases, we compute probabilities by *counting*
Basic Counting Principle

• Procedure (multiplication rule):
  - $r$ steps
  - $n_i$ choices at step $i$
  - Number of choices is $n_1 \times n_2 \times \cdots \times n_r$

• Example:
  - Number of license plates with 3 letters followed by 4 digits:
  - With no repetition (replacement), the number is:

• Example: Consider a set of objects $\{s_1, s_2, \ldots, s_n\}$. How many subsets of these objects are there (including the set itself and the empty set)?
  - Each object may or may not be in the subset (2 options)
  - The number of subsets is:

De Méré’s Paradox

• Counting can be very tricky

• Classic example: Throw three 6-sided dice. Is the probability that the sum of the outcomes is 11 equal to the probability that the sum of the outcomes is 12?

• De Méré’s argued that they are, since the number of different ways to obtain 11 and 12 are the same:
  - Sum=11: $\{6, 4, 1\}, \{6, 3, 2\}, \{5, 5, 1\}, \{5, 4, 2\}, \{5, 3, 3\}, \{4, 4, 3\}$
  - Sum=12: $\{6, 5, 1\}, \{6, 4, 2\}, \{6, 3, 3\}, \{5, 5, 2\}, \{5, 4, 3\}, \{4, 4, 4\}$

• This turned out to be false. Why?
Basic Types of Counting

- Assume we have \( n \) distinct objects \( a_1, a_2, \ldots, a_n \), e.g., digits, letters, etc.

- Some basic counting questions:
  - How many different ordered sequences of \( k \) objects can be formed out of the \( n \) objects with replacement?
  - How many different ordered sequences of \( k \leq n \) objects can be formed out of the \( n \) objects without replacement? (called \( k \)-permutations)
  - How many different unordered sequences (subsets) of \( k \leq n \) objects can be formed out of the \( n \) objects without replacement? (called \( k \)-combinations)
  - Given \( r \) nonnegative integers \( n_1, n_2, \ldots, n_r \) that sum to \( n \) (the number of objects), how many ways can the \( n \) objects be partitioned into \( r \) subsets (unordered sequences) with the \( i \)th subset having exactly \( n_i \) objects? (caller partitions, and is a generalization of combinations)

Ordered Sequences With and Without Replacement

- The number of ordered \( k \)-sequences from \( n \) objects with replacement is \( n \times n \times n \cdots \times n \) \( k \) times, i.e., \( n^k \)
  - Example: If \( n = 2 \), e.g., binary digits, the number of ordered \( k \)-sequences is \( 2^k \)
- The number of different ordered sequences of \( k \) objects that can be formed from \( n \) objects without replacement, i.e., the \( k \)-permutations, is
  \[
  n \times (n - 1) \times (n - 2) \cdots \times (n - k + 1)
  \]
  - If \( k = n \), the number is
  \[
  n \times (n - 1) \times (n - 2) \cdots \times 2 \times 1 = n! \text{ (n-factorial)}
  \]
  - Thus the number of \( k \)-permutations is: \( n!/(n-k)! \)
- Example: Consider the alphabet set \( \{A, B, C, D\} \), so \( n = 4 \)
  - The number of \( k = 2 \)-permutations of \( n = 4 \) objects is 12
  - They are: \( AB, AC, AD, BA, BC, BD, CA, CB, CD, DA, DB, \) and \( DC \)
Unordered Sequences Without Replacement

- Denote by \( \binom{n}{k} \) (\( n \) choose \( k \)) the number of unordered \( k \)-sequences that can be formed out of \( n \) objects without replacement, i.e., the \( k \)-combinations.

- Two different ways of constructing the \( k \)-permutations:
  1. Choose \( k \) objects (\( \binom{n}{k} \)), then order them (\( k! \) possible orders). This gives \( \binom{n}{k} \times k! \)
  2. Choose the \( k \) objects one at a time:
     \[ n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!} \text{ choices} \]

- Hence \( \binom{n}{k} \times k! = \frac{n!}{(n-k)!} \) or \( \binom{n}{k} = \frac{n!}{k!(n-k)!} \).

- Example: The number of \( k = 2 \) combinations of \{A, B, C, D\} is 6.
  They are AB, AC, AD, BC, BD, CD.

- What is the number of binary sequences of length \( n \) with exactly \( k \le n \) ones?

- Question: What is \( \sum_{k=0}^{n} \binom{n}{k} \)?

Finding Probability Using Counting

- Example: Die Rolling
  Roll a fair 6-sided die 6 times independently. Find the probability that outcomes from all six rolls are different.

  Solution:
  - \# outcomes yielding this event =
  - \# of points in sample space =
  - Probability =
Example: *The Birthday Problem*

$k$ people are selected at random. What is the probability that all $k$ birthdays will be different (neglect leap years)

Solution:

- Total # ways of assigning birthdays to $k$ people:

- # of ways of assigning birthdays to $k$ people with no two having the same birthday:

- Probability:

---

**Binomial Probabilities**

- Consider $n$ independent coin tosses, where $P(H) = p$ for $0 < p < 1$

- Outcome is a sequence of $H$s and $T$s of length $n$

\[
P(\text{sequence}) = p^{\#\text{heads}}(1 - p)^{\#\text{tails}}
\]

- The probability of $k$ heads in $n$ tosses is thus

\[
P(k \text{ heads}) = \sum_{\text{sequences with } k \text{ heads}} P(\text{sequence})
\]

\[
= \#(k-\text{head sequences}) \times p^k(1 - p)^{n-k}
\]

\[
= \binom{n}{k} p^k(1 - p)^{n-k}
\]

Check that it sums to 1
• Example: Toss a coin with bias $p$ independently 10 times.

Define the events $B = \{3$ out of 10 tosses are heads\}$ and $A = \{\text{first two tosses are heads}\}$. Find $P(A|B)$

Solution: The conditional probability is

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

All points in $B$ have the same probability $p^3(1-p)^7$, so we can find the conditional probability by counting:

◦ # points in $B$ beginning with two heads =

◦ # points in $B =$

◦ Probability =

---

Partitions

• Let $n_1, n_2, n_3, \ldots, n_r$ be such that

$$\sum_{i=1}^{r} n_i = n$$

How many ways can the $n$ objects be partitioned into $r$ subsets (unordered sequences) with the $i$th subset having exactly $n_i$ objects?

• If $r = 2$ with $n_1, n - n_1$ the answer is the $n_1$-combinations $\binom{n}{n_1}$

• Answer in general:
\[
\binom{n}{n_1, n_2, \ldots, n_r} = \binom{n}{n_1} \times \binom{n-n_1}{n_2} \times \binom{n-(n_1+n_2)}{n_3} \times \cdots \times \binom{n-\sum_{i=1}^{r-1} n_i}{n_r}
\]
\[
= \frac{n!}{n_1!(n-n_1)!} \times \frac{(n-n_1)!}{n_2!(n-(n_1+n_2))!} \times \cdots
\]
\[
= \frac{n!}{n_1!n_2! \cdots n_r!}
\]

- **Example: Balls and bins**

We have \(n\) balls and \(r\) bins. We throw each ball independently and at random into a bin. What is the probability that bin \(i = 1, 2, \ldots, r\) will have exactly \(n_i\) balls, where \(\sum_{i=1}^{r} n_i = n\)?

**Solution:**

- The probability of each outcome (sequence of \(n_i\) balls in bin \(i\)) is:

- \# of ways of partitioning the \(n\) balls into \(r\) bins such that bin \(i\) has exactly \(n_i\) balls is:

- Probability:
Example: *Cards*

Consider a perfectly shuffled 52-card deck dealt to 4 players. Find \( P(\text{each player gets an ace}) \)

Solution:

○ Size of the sample space is:

○ \# of ways of distributing the four aces:

○ \# of ways of dealing the remaining 48 cards:

○ Probability =
EE 178 Lecture Notes 1
Basic Probability

- Set Theory
- Elements of Probability
- Conditional probability
- Sequential Calculation of Probability
- Total Probability and Bayes Rule
- Independence
- Counting

Set Theory Basics

- A set is a collection of objects, which are its elements
  - $\omega \in A$ means that $\omega$ is an element of the set $A$
  - A set with no elements is called the empty set, denoted by $\emptyset$
- Types of sets:
  - Finite: $A = \{\omega_1, \omega_2, \ldots, \omega_n\}$
  - Countably infinite: $A = \{\omega_1, \omega_2, \ldots\}$, e.g., the set of integers
  - Uncountable: A set that takes a continuous set of values, e.g., the $[0, 1]$ interval, the real line, etc.
- A set can be described by all $\omega$ having a certain property, e.g., $A = [0, 1]$ can be written as $A = \{\omega : 0 \leq \omega \leq 1\}$
- A set $B \subset A$ means that every element of $B$ is an element of $A$
- A universal set $\Omega$ contains all objects of particular interest in a particular context, e.g., sample space for random experiment
Set Operations

- Assume a universal set \( \Omega \)

- Three basic operations:
  - Complementation: A complement of a set \( A \) with respect to \( \Omega \) is \( A^c = \{ \omega \in \Omega : \omega \notin A \} \), so \( \Omega^c = \emptyset \)
  - Intersection: \( A \cap B = \{ \omega : \omega \in A \text{ and } \omega \in B \} \)
  - Union: \( A \cup B = \{ \omega : \omega \in A \text{ or } \omega \in B \} \)

- Notation:
  - \( \bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \ldots \cup A_n \)
  - \( \bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \ldots \cap A_n \)

- A collection of sets \( A_1, A_2, \ldots, A_n \) are disjoint or mutually exclusive if \( A_i \cap A_j = \emptyset \) for all \( i \neq j \), i.e., no two of them have a common element

- A collection of sets \( A_1, A_2, \ldots, A_n \) partition \( \Omega \) if they are disjoint and \( \bigcup_{i=1}^{n} A_i = \Omega \)

- Venn Diagrams

(a) \( \Omega \)  
(b) \( A \)  
(c) \( B \)  
(d) \( B^c \)  
(e) \( A \cap B \)  
(f) \( A \cup B \)
Algebra of Sets

- Basic relations:
  1. \( S \cap \Omega = S \)
  2. \((A^c)^c = A\)
  3. \(A \cap A^c = \emptyset\)
  4. Commutative law: \( A \cup B = B \cup A \)
  5. Associative law: \( A \cup (B \cup C) = (A \cup B) \cup C \)
  6. Distributive law: \( A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \)
  7. DeMorgan’s law: \((A \cap B)^c = A^c \cup B^c\)
    - DeMorgan’s law can be generalized to \( n \) events:
      \[
      (\bigcap_{i=1}^{n} A_i)^c = \bigcup_{i=1}^{n} A_i^c
      \]

- These can all be proven using the definition of set operations or visualized using Venn Diagrams

Elements of Probability

- Probability theory provides the mathematical rules for assigning probabilities to outcomes of random experiments, e.g., coin flips, packet arrivals, noise voltage

- Basic elements of probability:
  - Sample space: The set of all possible “elementary” or “finest grain” outcomes of the random experiment (also called sample points)
    - The sample points are all disjoint
    - The sample points are collectively exhaustive, i.e., together they make up the entire sample space
  - Events: Subsets of the sample space
  - Probability law: An assignment of probabilities to events in a mathematically consistent way
Discrete Sample Spaces

- Sample space is called discrete if it contains a countable number of sample points.

- Examples:
  - Flip a coin once: $\Omega = \{H, T\}$
  - Flip a coin three times: $\Omega = \{HHH, HHT, HTH, \ldots\} = \{H, T\}^3$
  - Flip a coin $n$ times: $\Omega = \{H, T\}^n$ (set of sequences of H and T of length $n$)
  - Roll a die once: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Roll a die twice: $\Omega = \{(1, 1), (1, 2), (2, 1), \ldots, (6, 6)\} = \{1, 2, 3, 4, 5, 6\}^2$
  - Flip a coin until the first heads appears: $\Omega = \{H, TH, TTH, TTTH, \ldots\}$
  - Number of packets arriving in time interval $(0, T]$ at a node in a communication network: $\Omega = \{0, 1, 2, 3, \ldots\}$

Note that the first five examples have finite $\Omega$, whereas the last two have countably infinite $\Omega$. Both types are called discrete.

- Sequential models: For sequential experiments, the sample space can be described in terms of a tree, where each outcome corresponds to a terminal node (or a leaf).

Example: Three flips of a coin
• Example: Roll a fair four-sided die twice.
  Sample space can be represented by a tree as above, or graphically

\begin{figure}
\centering
\includegraphics[width=0.5\textwidth]{dice_tree.png}
\end{figure}

Continuous Sample Spaces

• A continuous sample space consists of a continuum of points and thus contains an uncountable number of points

• Examples:
  ○ Random number between 0 and 1: \( \Omega = [0, 1] \)
  ○ Suppose we pick two numbers at random between 0 and 1, then the sample space consists of all points in the unit square, i.e., \( \Omega = [0, 1]^2 \)
o Packet arrival time: \( t \in (0, \infty) \), thus \( \Omega = (0, \infty) \)

o Arrival times for \( n \) packets: \( t_i \in (0, \infty) \), for \( i = 1, 2, \ldots, n \), thus \( \Omega = (0, \infty)^n \)

- A sample space is said to be *mixed* if it is neither discrete nor continuous, e.g., \( \Omega = [0, 1] \cup \{3\} \)

### Events

- Events are *subsets* of the sample space. An event occurs if the outcome of the experiment belongs to the event

- Examples:
  - Any outcome (sample point) is an event (also called an elementary event), e.g., \{HTH\} in three coin flips experiment or \{0.35\} in the picking of a random number between 0 and 1 experiment
  - Flip coin 3 times and get exactly one H. This is a more complicated event, consisting of three sample points \{TTH, THT, HTT\}
  - Flip coin 3 times and get an odd number of H’s. The event is \{TTH, THT, HTT, HHH\}
  - Pick a random number between 0 and 1 and get a number between 0.0 and 0.5. The event is \[0, 0.5\]

- An event might have *no points* in it, i.e., be the empty set \( \emptyset \)
Axioms of Probability

• Probability law (measure or function) is an assignment of probabilities to events (subsets of sample space $\Omega$) such that the following three axioms are satisfied:

1. $P(A) \geq 0$, for all $A$ (nonnegativity)
2. $P(\Omega) = 1$ (normalization)
3. If $A$ and $B$ are disjoint ($A \cap B = \emptyset$), then

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

(additivity)

More generally,

3’. If the sample space has an infinite number of points and $A_1, A_2, \ldots$ are disjoint events, then

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

• Mimics relative frequency, i.e., perform the experiment $n$ times (e.g., roll a die $n$ times). If the number of occurrences of $A$ is $n_A$, define the relative frequency of an event $A$ as $f_A = n_A/n$.

  ○ Probabilities are nonnegative (like relative frequencies)
  ○ Probability something happens is 1 (again like relative frequencies)
  ○ Probabilities of disjoint events add (again like relative frequencies)

• Analogy: Except for normalization, probability is a measure much like

  ○ mass
  ○ length
  ○ area
  ○ volume

They all satisfy axioms 1 and 3.

This analogy provides some intuition but is not sufficient to fully understand probability theory — other aspects such as conditioning, independence, etc., are unique to probability.
• Recall that sample space $\Omega$ is said to be \textit{discrete} if it is countable.

• The probability measure $P$ can be simply defined by first assigning probabilities to outcomes, i.e., elementary events $\{\omega\}$, such that:

\[ P(\{\omega\}) \geq 0, \text{ for all } \omega \in \Omega, \text{ and} \]
\[ \sum_{\omega \in \Omega} P(\{\omega\}) = 1 \]

• The probability of any other event $A$ (by the additivity axiom) is simply

\[ P(A) = \sum_{\omega \in A} P(\{\omega\}) \]

• Examples:
  
  ◦ For the coin flipping experiment, assign

  \[ P(\{H\}) = p \text{ and } P(\{T\}) = 1 - p, \text{ for } 0 \leq p \leq 1 \]
  
  Note: $p$ is the \textit{bias} of the coin, and a coin is \textit{fair} if $p = \frac{1}{2}$.

  ◦ For the die rolling experiment, assign

  \[ P(\{i\}) = \frac{1}{6}, \text{ for } i = 1, 2, \ldots, 6 \]

  The probability of the event “the outcome is even”, $A = \{2, 4, 6\}$, is

  \[ P(A) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{2} \]
If $\Omega$ is countably infinite, we can again assign probabilities to elementary events.

Example: Assume $\Omega = \{1, 2, \ldots\}$, assign probability $2^{-k}$ to event $\{k\}$

The probability of the event “the outcome is even”

\[
P(\text{outcome is even}) = P(\{2, 4, 6, 8, \ldots\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) + \ldots\]

\[
= \sum_{k=1}^{\infty} P(\{2k\}) = \sum_{k=1}^{\infty} 2^{-2k} = \frac{1}{3}
\]

---

**Probability for Continuous Sample Space**

- Recall that if a sample space is *continuous*, $\Omega$ is uncountably infinite.
- For continuous $\Omega$, we cannot in general define the probability measure $P$ by first assigning probabilities to outcomes.
- To see why, consider assigning a uniform probability measure to $\Omega = (0, 1]$
  - In this case the probability of each single outcome event is zero.
  - How do we find the probability of an event such as $A = \left[\frac{1}{2}, \frac{3}{4}\right]$?
- For this example we can define uniform probability measure over $[0, 1]$ by assigning to an event $A$, the probability

\[
P(A) = \text{length of } A,
\]

E.g., $P([0, 1/3] \cup [2/3, 1]) = 2/3$

Check that this is a legitimate assignment.
Another example: Romeo and Juliet have a date. Each arrives late with a random delay of up to 1 hour. Each will wait only 1/4 of an hour before leaving. What is the probability that Romeo and Juliet will meet?

Solution: The pair of delays is equivalent to that achievable by picking two random numbers between 0 and 1. Define probability of an event as its area

The event of interest is represented by the cross hatched region

\[ \text{Probability of the event is:} \]
\[ \text{area of crosshatched region} = 1 - 2 \times \frac{1}{2}(0.75)^2 = 0.4375 \]

Basic Properties of Probability

- There are several useful properties that can be derived from the axioms of probability:
  1. \( P(A^c) = 1 - P(A) \)
     - \( P(\emptyset) = 0 \)
     - \( P(A) \leq 1 \)
  2. If \( A \subseteq B \), then \( P(A) \leq P(B) \)
  3. \( P(A \cup B) = P(A) + P(B) - P(A \cap B) \)
  4. \( P(A \cup B) \leq P(A) + P(B) \), or in general
     \[ P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i) \]

  This is called the Union of Events Bound

- These properties can be proved using the axioms of probability and visualized using Venn diagrams
Conditional Probability

- Conditional probability allows us to compute probabilities of events based on partial knowledge of the outcome of a random experiment.

- Examples:
  - We are told that the sum of the outcomes from rolling a die twice is 9. What is the probability the outcome of the first die was a 6?
  - A spot shows up on a radar screen. What is the probability that there is an aircraft?
  - You receive a 0 at the output of a digital communication system. What is the probability that a 0 was sent?

- As we shall see, conditional probability provides us with two methods for computing probabilities of events: the **sequential** method and the **divide-and-conquer** method.

- It is also the basis of inference in statistics: make an observation and reason about the cause.

- In general, given an event \( B \) has occurred, we wish to find the probability of another event \( A \), \( P(A|B) \).

- If all elementary outcomes are equally likely, then

  \[
P(A|B) = \frac{\text{# of outcomes in both } A \text{ and } B}{\text{# of outcomes in } B}
\]

- In general, if \( B \) is an event such that \( P(B) \neq 0 \), the **conditional probability** of any event \( A \) given \( B \) is defined as

  \[
P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ or } \frac{P(A, B)}{P(B)}
\]

- The function \( P(\cdot|B) \) for fixed \( B \) specifies a probability law, i.e., it satisfies the axioms of probability.
Example

• Roll a fair four-sided die twice. So, the sample space is \( \{1, 2, 3, 4\}^2 \). All sample points have probability \( \frac{1}{16} \).

Let \( B \) be the event that the minimum of the two die rolls is 2 and \( A_m \), for \( m = 1, 2, 3, 4 \), be the event that the maximum of the two die rolls is \( m \). Find \( P(A_m|B) \).

• Solution:

![Diagram showing a 4x4 grid with labels for first and second rolls.]

Conditional Probability Models

• Before: Probability law \( \Rightarrow \) conditional probabilities

• Reverse is often more natural: Conditional probabilities \( \Rightarrow \) probability law

• We use the chain rule (also called multiplication rule):

By the definition of conditional probability, \( P(A \cap B) = P(A|B)P(B) \). Suppose that \( A_1, A_2, \ldots, A_n \) are events, then
\[ P(A_1 \cap A_2 \cap A_3 \cdots \cap A_n) \]
\[ = P(A_1 \cap A_2 \cap A_3 \cdots \cap A_{n-1}) \times P(A_n \mid A_1 \cap A_2 \cap A_3 \cdots \cap A_{n-1}) \]
\[ = P(A_1 \cap A_2 \cap A_3 \cdots \cap A_{n-2}) \times P(A_{n-1} \mid A_1 \cap A_2 \cap A_3 \cdots \cap A_{n-2}) \]
\[ \times P(A_n \mid A_1 \cap A_2 \cap A_3 \cdots \cap A_{n-1}) \]
\[ \vdots \]
\[ = P(A_1) \times P(A_2 \mid A_1) \times P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap A_2 \cap A_3 \cdots \cap A_{n-1}) \]
\[ = \prod_{i=1}^{n} P(A_i \mid A_1, A_2, \ldots, A_{i-1}), \]

where \( A_0 = \emptyset \)

---

**Sequential Calculation of Probabilities**

- **Procedure:**
  1. Construct a tree description of the sample space for a sequential experiment.
  2. Assign the conditional probabilities on the corresponding branches of the tree.
  3. By the chain rule, the probability of an outcome can be obtained by multiplying the conditional probabilities along the path from the root to the leaf node corresponding to the outcome.

- **Example (Radar Detection):** Let \( A \) be the event that an aircraft is flying above and \( B \) be the event that the radar detects it. Assume \( P(A) = 0.05 \), \( P(B \mid A) = 0.99 \), and \( P(B \mid A^c) = 0.1 \).

  What is the probability of
  - Missed detection?, i.e., \( P(A \cap B^c) \)
  - False alarm?, i.e., \( P(B \cap A^c) \)

  The sample space is: \( \Omega = \{(A, B), (A^c, B), (A, B^c), (A^c, B^c)\} \)
Solution: Represent the sample space by a tree with conditional probabilities on its edges

Example: Three cards are drawn at random (without replacement) from a deck of cards. What is the probability of not drawing a heart?

Solution: Let $A_i$, $i = 1, 2, 3$, represent the event of no heart in the $i$th draw. We can represent the sample space as:

$$\Omega = \{(A_1, A_2, A_3), (A_1^c, A_2, A_3), \ldots, (A_1^c, A_2^c, A_3^c)\}$$

To find the probability law, we represent the sample space by a tree, write conditional probabilities on branches, and use the chain rule.
Total probability – Divide and Conquer Method

- Let $A_1, A_2, \ldots, A_n$ be events that partition $\Omega$, i.e., that are disjoint ($A_i \cap A_j = \emptyset$ for $i \neq j$) and $\bigcup_{i=1}^{n} A_i = \Omega$. Then for any event $B$

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

This is called the Law of Total Probability. It also holds for $n = \infty$. It allows us to compute the probability of a complicated event from knowledge of probabilities of simpler events.

- Example: Chess tournament, 3 types of opponents for a certain player.
  - $P(\text{Type 1}) = 0.5$, $P(\text{Win} | \text{Type1}) = 0.3$
  - $P(\text{Type 2}) = 0.25$, $P(\text{Win} | \text{Type2}) = 0.4$
  - $P(\text{Type 3}) = 0.25$, $P(\text{Win} | \text{Type3}) = 0.5$

What is probability of player winning?
Solution: Let $B$ be the event of winning and $A_i$ be the event of playing Type $i$, $i = 1, 2, 3$:

$$P(B) = \sum_{i=1}^{3} P(A_i)P(B|A_i)$$

$$= 0.5 \times 0.3 + 0.25 \times 0.4 + 0.25 \times 0.5$$

$$= 0.375$$

Bayes Rule

- Let $A_1, A_2, \ldots, A_n$ be nonzero probability events (the causes) that partition $\Omega$, and let $B$ be a nonzero probability event (the effect)

- We often know the \textit{a priori} probabilities $P(A_i)$, $i = 1, 2, \ldots, n$ and the conditional probabilities $P(B|A_i)$s and wish to find the \textit{a posteriori probabilities} $P(A_j|B)$ for $j = 1, 2, \ldots, n$

- From the definition of conditional probability, we know that

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

By the law of total probability

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$
Substituting we obtain Bayes rule

\[ P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)} \text{ for } j = 1, 2, \ldots, n \]

- Bayes rule also applies when the number of events \( n = \infty \)
- Radar Example: Recall that \( A \) is event that the aircraft is flying above and \( B \) is the event that the aircraft is detected by the radar. What is the probability that an aircraft is actually there given that the radar indicates a detection?
  Recall \( P(A) = 0.05, P(B|A) = 0.99, P(B|A^c) = 0.1 \). Using Bayes rule:

\[
P(\text{there is an aircraft|radar detects it}) = P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{0.05 \times 0.99}{0.05 \times 0.99 + 0.95 \times 0.1} = 0.3426
\]

**Binary Communication Channel**

- Consider a noisy binary communication channel, where 0 or 1 is sent and 0 or 1 is received. Assume that 0 is sent with probability 0.2 (and 1 is sent with probability 0.8)

  The channel is noisy. If a 0 is sent, a 0 is received with probability 0.9, and if a 1 is sent, a 1 is received with probability 0.975

- We can represent this channel model by a probability transition diagram

\[
P(\{0\}) = 0.2 \quad \begin{array}{c}
0 \\
0.1
\end{array}
\quad
\begin{array}{c}
0.9 \\
0.025
\end{array}
\quad
P(\{1\}) = 0.8 \quad \begin{array}{c}
1 \\
1
\end{array}
\quad
P(\{\{0\}|\{0\}) = 0.9
\]

Given that 0 is received, find the probability that 0 was sent
This is a random experiment with sample space \( \Omega = \{(0, 0), (0, 1), (1, 0), (1, 1)\} \), where the first entry is the bit sent and the second is the bit received.

Define the two events
\[
A = \{0 \text{ is sent}\} = \{(0, 1), (0, 0)\}, \quad \text{and} \quad B = \{0 \text{ is received}\} = \{(0, 0), (1, 0)\}
\]

The probability measure is defined via the \( P(A) \), \( P(B|A) \), and \( P(B^c|A^c) \) provided on the probability transition diagram of the channel.

To find \( P(A|B) \), the *a posteriori* probability that a 0 was sent. We use Bayes rule
\[
P(A|B) = \frac{P(B|A)P(A)}{P(A)P(B|A) + P(A^c)P(B|A^c)},
\]
which is much higher than the *a priori* probability of \( A (= 0.2) \).

---

**Independence**

It often happens that the knowledge that a certain event \( B \) has occurred has no effect on the probability that another event \( A \) has occurred, i.e.,
\[
P(A|B) = P(A)
\]
In this case we say that the two events are statistically independent.

Equivalently, two events are said to be *statistically independent* if
\[
P(A, B) = P(A)P(B)
\]
So, in this case, \( P(A|B) = P(A) \) and \( P(B|A) = P(B) \).

Example: Assuming that the binary channel of the previous example is used to send two bits independently, what is the probability that both bits are in error?
Solution:

- Define the two events

\[ E_1 = \{ \text{First bit is in error} \} \]
\[ E_2 = \{ \text{Second bit is in error} \} \]

- Since the bits are sent independently, the probability that both are in error is

\[ P(E_1, E_2) = P(E_1)P(E_2) \]

Also by symmetry, \( P(E_1) = P(E_2) \)

To find \( P(E_1) \), we express \( E_1 \) in terms of the events \( A \) and \( B \) as

\[ E_1 = (A_1 \cap B_1^c) \cup (A_1^c \cap B_1), \]

- Thus,

\[ P(E_1) = P(A_1, B_1^c) + P(A_1^c, B_1) \]
\[ = P(A_1)P(B_1^c|A_1) + P(A_1^c)P(B_1|A_1^c) \]
\[ = 0.2 \times 0.1 + 0.8 \times 0.025 = 0.04 \]

- The probability that the two bits are in error

\[ P(E_1, E_2) = (0.04)^2 = 16 \times 10^{-4} \]

- In general \( A_1, A_2, \ldots, A_n \) are mutually independent if for each subset of the events \( A_{i_1}, A_{i_2}, \ldots, A_{i_k} \)

\[ P(A_{i_1}, A_{i_2}, \ldots, A_{i_k}) = \prod_{j=1}^{k} P(A_{i_j}) \]

- Note: \( P(A_1, A_2, \ldots, A_n) = \prod_{j=1}^{n} P(A_i) \) alone is not sufficient for independence.
Example: Roll two fair dice independently. Define the events

- \( A = \{ \text{First die } = 1, 2, \text{ or } 3 \} \)
- \( B = \{ \text{First die } = 2, 3, \text{ or } 6 \} \)
- \( C = \{ \text{Sum of outcomes } = 9 \} \)

Are \( A, B, \) and \( C \) independent?

Solution:
Since the dice are fair and the experiments are performed independently, the probability of any pair of outcomes is \( \frac{1}{36} \), and

\[
P(A) = \frac{1}{2}, \quad P(B) = \frac{1}{2}, \quad P(C) = \frac{1}{9}
\]

Since \( A \cap B \cap C = \{(3, 6)\} \), \( P(A, B, C) = \frac{1}{36} = P(A)P(B)P(C) \)

But are \( A, B, \) and \( C \) independent? Let’s find

\[
P(A, B) = \frac{2}{6} = \frac{1}{3} \neq \frac{1}{4} = P(A)P(B),
\]

and thus \( A, B, \) and \( C \) are not independent!
• Also, independence of subsets does not necessarily imply independence

Example: Flip a fair coin twice independently. Define the events:
- $A$: First toss is Head
- $B$: Second toss is Head
- $C$: First and second toss have different outcomes

$A$ and $B$ are independent, $A$ and $C$ are independent, and $B$ and $C$ are independent

Are $A, B, C$ mutually independent?

Clearly not, since if you know $A$ and $B$, you know that $C$ could not have occurred, i.e., $P(A, B, C) = 0 \neq P(A)P(B)P(C) = 1/8$

---

Counting

• Discrete uniform law:
  - Finite sample space where all sample points are equally probable:
    $$P(A) = \frac{\text{number of sample points in } A}{\text{total number of sample points}}$$
  - Variation: all outcomes in $A$ are equally likely, each with probability $p$. Then,
    $$P(A) = p \times (\text{number of elements of } A)$$

• In both cases, we compute probabilities by counting
Basic Counting Principle

- Procedure (multiplication rule):
  - $r$ steps
  - $n_i$ choices at step $i$
  - Number of choices is $n_1 \times n_2 \times \cdots \times n_r$

- Example:
  - Number of license plates with 3 letters followed by 4 digits:
  - With no repetition (replacement), the number is:

- Example: Consider a set of objects $\{s_1, s_2, \ldots, s_n\}$. How many subsets of these objects are there (including the set itself and the empty set)?
  - Each object may or may not be in the subset (2 options)
  - The number of subsets is:

De Méré’s Paradox

- Counting can be very tricky

- Classic example: Throw three 6-sided dice. Is the probability that the sum of the outcomes is 11 equal to the probability that the sum of the outcomes is 12?

- De Méré’s argued that they are, since the number of different ways to obtain 11 and 12 are the same:
  - Sum=11: $\{6, 4, 1\}$, $\{6, 3, 2\}$, $\{5, 5, 1\}$, $\{5, 4, 2\}$, $\{5, 3, 3\}$, $\{4, 4, 3\}$
  - Sum=12: $\{6, 5, 1\}$, $\{6, 4, 2\}$, $\{6, 3, 3\}$, $\{5, 5, 2\}$, $\{5, 4, 3\}$, $\{4, 4, 4\}$

- This turned out to be false. Why?
Basic Types of Counting

- Assume we have \( n \) distinct objects \( a_1, a_2, \ldots, a_n \), e.g., digits, letters, etc.

- Some basic counting questions:
  - How many different ordered sequences of \( k \) objects can be formed out of the \( n \) objects with replacement?
  - How many different ordered sequences of \( k \leq n \) objects can be formed out of the \( n \) objects without replacement? (called \( k \)-permutations)
  - How many different unordered sequences (subsets) of \( k \leq n \) objects can be formed out of the \( n \) objects without replacement? (called \( k \)-combinations)
  - Given \( r \) nonnegative integers \( n_1, n_2, \ldots, n_r \) that sum to \( n \) (the number of objects), how many ways can the \( n \) objects be partitioned into \( r \) subsets (unordered sequences) with the \( i \)th subset having exactly \( n_i \) objects? (called \( \text{partitions} \), and is a generalization of combinations)

Ordered Sequences With and Without Replacement

- The number of ordered \( k \)-sequences from \( n \) objects with replacement is \( n \times n \times n \ldots \times n \) \( k \) times, i.e., \( n^k \)
  - Example: If \( n = 2 \), e.g., binary digits, the number of ordered \( k \)-sequences is \( 2^k \)

- The number of different ordered sequences of \( k \) objects that can be formed from \( n \) objects without replacement, i.e., the \( k \)-permutations, is
  \[
  n \times (n - 1) \times (n - 2) \cdots \times (n - k + 1)
  \]
  If \( k = n \), the number is
  \[
  n \times (n - 1) \times (n - 2) \cdots \times 2 \times 1 = n! \text{ (n-factorial)}
  \]
  Thus the number of \( k \)-permutations is: \( n!/(n - k)! \)

- Example: Consider the alphabet set \( \{A, B, C, D\} \), so \( n = 4 \)
  The number of \( k = 2 \)-permutations of \( n = 4 \) objects is 12
  They are: \( AB, AC, AD, BA, BC, BD, CA, CB, CD, DA, DB, \) and \( DC \)
Unordered Sequences Without Replacement

• Denote by \( \binom{n}{k} \) (\( n \) choose \( k \)) the number of unordered \( k \)-sequences that can be formed out of \( n \) objects without replacement, i.e., the \( k \)-combinations

• Two different ways of constructing the \( k \)-permutations:
  1. Choose \( k \) objects (\( \binom{n}{k} \)), then order them (\( k! \) possible orders). This gives \( \binom{n}{k} \times k! \)
  2. Choose the \( k \) objects one at a time:
      \[
      n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!} \]
      \( \text{choices} \)

• Hence \( \binom{n}{k} \times k! = \frac{n!}{(n-k)!} \) or \( \binom{n}{k} = \frac{n!}{k!(n-k)!} \)

• Example: The number of \( k = 2 \)-combinations of \{A, B, C, D\} is 6

  They are AB, AC, AD, BC, BD, CD

• What is the number of binary sequences of length \( n \) with exactly \( k \leq n \) ones?

• Question: What is \( \sum_{k=0}^{n} \binom{n}{k} \)?

Finding Probability Using Counting

• Example: Die Rolling

Roll a fair 6-sided die 6 times independently. Find the probability that outcomes from all six rolls are different

Solution:

○ # outcomes yielding this event =

○ # of points in sample space =

○ Probability =
Example: *The Birthday Problem*

$k$ people are selected at random. What is the probability that all $k$ birthdays will be different (neglect leap years)

Solution:

- Total # ways of assigning birthdays to $k$ people:

- # of ways of assigning birthdays to $k$ people with no two having the same birthday:

- Probability:

---

**Binomial Probabilities**

- Consider $n$ independent coin tosses, where $P(H) = p$ for $0 < p < 1$

- Outcome is a sequence of $H$s and $T$s of length $n$

\[
P(\text{sequence}) = p^{\#\text{heads}}(1-p)^{\#\text{tails}}
\]

- The probability of $k$ heads in $n$ tosses is thus

\[
P(k \text{ heads}) = \sum_{\text{sequences with } k \text{ heads}} P(\text{sequence})
\]

\[
= \#(k-\text{head sequences}) \times p^k(1-p)^{n-k}
\]

\[
= \binom{n}{k} p^k(1-p)^{n-k}
\]

Check that it sums to 1
• Example: Toss a coin with bias $p$ independently 10 times. Define the events $B = \{3$ out of 10 tosses are heads\} and $A = \{first$ two tosses are heads\}. Find $P(A|B)$

Solution: The conditional probability is

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

All points in $B$ have the same probability $p^3(1-p)^7$, so we can find the conditional probability by counting:

- # points in $B$ beginning with two heads =
- # points in $B =$
- Probability =

EE 178: Basic Probability

---

### Partitions

- Let $n_1, n_2, n_3, \ldots, n_r$ be such that

$$\sum_{i=1}^{r} n_i = n$$

How many ways can the $n$ objects be partitioned into $r$ subsets (unordered sequences) with the $i$th subset having exactly $n_i$ objects?

- If $r = 2$ with $n_1, n - n_1$ the answer is the $n_1$-combinations $\binom{n}{n_1}$

- Answer in general:
\[
\binom{n}{n_1 \ n_2 \ldots n_r} = \binom{n}{n_1} \times \binom{n-n_1}{n_2} \times \binom{n-(n_1+n_2)}{n_3} \times \ldots \times \binom{n-\sum_{i=1}^{r-1} n_i}{n_r}
\]
\[
= \frac{n!}{n_1!(n-n_1)!} \times \frac{(n-n_1)!}{n_2!(n-(n_1+n_2))!} \times \ldots \times \frac{n_i!}{n_r!}
\]

- **Example: Balls and bins**

We have \( n \) balls and \( r \) bins. We throw each ball independently and at random into a bin. What is the probability that bin \( i = 1, 2, \ldots, r \) will have exactly \( n_i \) balls, where \( \sum_{i=1}^{r} n_i = n \)?

**Solution:**

- The probability of each outcome (sequence of \( n_i \) balls in bin \( i \)) is:

- \# of ways of partitioning the \( n \) balls into \( r \) bins such that bin \( i \) has exactly \( n_i \) balls is:

- Probability:
Example: *Cards*

Consider a perfectly shuffled 52-card deck dealt to 4 players. Find $P(\text{each player gets an ace})$

Solution:

- Size of the sample space is:

- # of ways of distributing the four aces:

- # of ways of dealing the remaining 48 cards:

- Probability =
• Definition

• Discrete Random Variables: Probability mass function (pmf)

• Continuous Random Variables: Probability density function (pdf)

• Mean and Variance

• Cumulative Distribution Function (cdf)

• Functions of Random Variables


Random Variable

• A random variable is a real-valued variable that takes on values randomly.
  Sounds nice, but not terribly precise or useful

• Mathematically, a random variable (r.v.) $X$ is a real-valued function $X(\omega)$ over
  the sample space $\Omega$ of a random experiment, i.e., $X : \Omega \to \mathbb{R}$

• Randomness comes from the fact that outcomes are random ($X(\omega)$ is a
  deterministic function)

• Notations:
  ◦ Always use upper case letters for random variables ($X, Y, \ldots$)
  ◦ Always use lower case letters for values of random variables: $X = x$ means
    that the random variable $X$ takes on the value $x$
• Examples:

1. Flip a coin \( n \) times. Here \( \Omega = \{H, T\}^n \). Define the random variable \( X \in \{0, 1, 2, \ldots, n\} \) to be the number of heads.

2. Roll a 4-sided die twice.
   (a) Define the random variable \( X \) as the maximum of the two rolls.
   (b) Define the random variable \( Y \) to be the sum of the outcomes of the two rolls.
   (c) Define the random variable \( Z \) to be 0 if the sum of the two rolls is odd and 1 if it is even.

3. Flip coin until first heads shows up. Define the random variable \( X \in \{1, 2, \ldots\} \) to be the number of flips until the first heads.

4. Let \( \Omega = \mathbb{R} \). Define the two random variables
   (a) \( X = \omega \)
   (b) \( Y = \begin{cases} +1 & \text{for } \omega \geq 0 \\ -1 & \text{otherwise} \end{cases} \)

5. \( n \) packets arrive at a node in a communication network. Here \( \Omega \) is the set of arrival time sequences \( (t_1, t_2, \ldots, t_n) \in (0, \infty)^n \).
   (a) Define the random variable \( N \) to be the number of packets arriving in the interval \( (0, 1] \).
   (b) Define the random variable \( T \) to be the first interarrival time.
Why do we need random variables?

- Random variable can represent the gain or loss in a random experiment, e.g., stock market
- Random variable can represent a measurement over a random experiment, e.g., noise voltage on a resistor

- In most applications we care more about these costs/measurements than the underlying probability space
- Very often we work directly with random variables without knowing (or caring to know) the underlying probability space

Specifying a Random Variable

- Specifying a random variable means being able to determine the probability that $X \in A$ for any event $A \subset \mathbb{R}$, e.g., any interval

- To do so, we consider the inverse image of the set $A$ under $X(\omega)$, $\{w : X(\omega) \in A\}$

- So, $X \in A$ iff $\omega \in \{w : X(\omega) \in A\}$, thus $P(\{X \in A\}) = P(\{w : X(\omega) \in A\})$, or in short

$$P\{X \in A\} = P\{w : X(\omega) \in A\}$$
• Example: Roll fair 4-sided die twice independently: Define the r.v. $X$ to be the maximum of the two rolls. What is the $P\{0.5 < X \leq 2\}$?

- We classify r.v.s as:
  - Discrete: $X$ can assume only one of a countable number of values. Such r.v. can be specified by a Probability Mass Function (pmf). Examples 1, 2, 3, 4(b), and 5(a) are of discrete r.v.s
  - Continuous: $X$ can assume one of a continuum of values and the probability of each value is 0. Such r.v. can be specified by a Probability Density Function (pdf). Examples 4(a) and 5(b) are of continuous r.v.s
  - Mixed: $X$ is neither discrete nor continuous. Such r.v. (as well as discrete and continuous r.v.s) can be specified by a Cumulative Distribution Function (cdf)
Discrete Random Variables

- A random variable is said to be discrete if for some countable set $\mathcal{X} \subset \mathbb{R}$, i.e., $\mathcal{X} = \{x_1, x_2, \ldots\}$, $P\{X \in \mathcal{X}\} = 1$

- Examples 1, 2, 3, 4(b), and 5(a) are discrete random variables

- Here $X(\omega)$ partitions $\Omega$ into the sets $\{\omega : X(\omega) = x_i\}$ for $i = 1, 2, \ldots$. Therefore, to specify $X$, it suffices to know $P\{X = x_i\}$ for all $i$

- A discrete random variable is thus completely specified by its probability mass function (pmf)

$$p_X(x) = P\{X = x\} \text{ for all } x \in \mathcal{X}$$

- Clearly $p_X(x) \geq 0$ and $\sum_{x \in \mathcal{X}} p_X(x) = 1$

- Example: Roll a fair 4-sided die twice independently: Define the r.v. $X$ to be the maximum of the two rolls

$p_X(x)$:
• Note that $p_X(x)$ can be viewed as a probability measure over a discrete sample space (even though the original sample space may be continuous as in examples 4(b) and 5(a))

• The probability of any event $A \subset R$ is given by

$$P\{X \in A\} = \sum_{x \in A \cap X} p_X(x)$$

For the previous example $P\{1 < X \leq 2.5 \text{ or } X \geq 3.5\} =$

• Notation: We use $X \sim p_X(x)$ or simply $X \sim p(x)$ to mean that the discrete random variable $X$ has pmf $p_X(x)$ or $p(x)$

Famous Probability Mass Functions

• **Bernoulli**: $X \sim Bern(p)$ for $0 \leq p \leq 1$ has pmf

$$p_X(1) = p, \text{ and } p_X(0) = 1 - p$$

• **Geometric**: $X \sim Geom(p)$ for $0 \leq p \leq 1$ has pmf

$$p_X(k) = p(1 - p)^{k-1} \text{ for } k = 1, 2, \ldots$$

This r.v. represents, for example, the number of coin flips until the first heads shows up (assuming independent coin flips)
• **Binomial:** $X \sim B(n, p)$ for integer $n > 0$ and $0 \leq p \leq 1$ has pmf

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k = 0, 1, 2, \ldots, n$$

The maximum of $p_X(k)$ is attained at

$$k^* = \begin{cases} (n+1)p, & \text{if } (n+1)p \text{ is an integer} \\ \lfloor (n+1)p \rfloor, & \text{otherwise} \end{cases}$$

The binomial r.v. represents, for example, the number of heads in $n$ independent coin flips (see page 48 of Lecture Notes 1)

• **Poisson:** $X \sim \text{Poisson}(\lambda)$ for $\lambda > 0$ (called the rate) has pmf

$$p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda} \text{ for } k = 0, 1, 2, \ldots$$

The maximum of $p_X(k)$ attained at

$$k^* = \begin{cases} \lambda, & \text{if } \lambda \text{ is an integer} \\ \lfloor \lambda \rfloor, & \text{otherwise} \end{cases}$$

The Poisson r.v. often represents the number of random events, e.g., arrivals of packets, photons, customers, etc. in some time interval, e.g., $(0, 1]$
• Poisson is the limit of Binomial when \( p \propto \frac{1}{n} \), as \( n \to \infty \)

To show this let \( X_n \sim B(n, \frac{\lambda}{n}) \) for \( \lambda > 0 \). For any fixed nonnegative integer \( k \):

\[
p_{X_n}(k) = \binom{n}{k} \left( \frac{\lambda}{n} \right)^k \left( 1 - \frac{\lambda}{n} \right)^{(n-k)}
= \frac{n(n-1)\ldots(n-k+1)}{k!} \frac{\lambda^k}{n^k} \left( \frac{n-\lambda}{n} \right)^{n-k}
= \frac{n(n-1)\ldots(n-k+1)}{(n-\lambda)^k} \frac{\lambda^k}{k!} \left( \frac{n-\lambda}{n} \right)^n
\to \frac{\lambda^k}{k!} e^{-\lambda} \text{ as } n \to \infty
\]

Continuous Random Variables

• Suppose a r.v. \( X \) can take on a continuum of values each with probability 0

Examples:
  - Pick a number between 0 and 1
  - Measure the voltage across a heated resistor
  - Measure the phase of a random sinusoid . . .

• How do we describe probabilities of interesting events?

• Idea: For discrete r.v., we sum a pmf over points in a set to find its probability.
  For continuous r.v., integrate a probability density over a set to find its probability — analogous to mass density in physics (integrate mass density to get the mass)
Probability Density Function

- A continuous r.v. $X$ can be specified by a *probability density function* $f_X(x)$ (pdf) such that, for any event $A$,

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

  For example, for $A = (a, b]$, the probability can be computed as

  $$P\{X \in (a, b]\} = \int_a^b f_X(x) \, dx$$

- Properties of $f_X(x)$:
  1. $f_X(x) \geq 0$
  2. $\int_{-\infty}^{\infty} f_X(x) \, dx = 1$

- Important note: $f_X(x)$ should not be interpreted as the probability that $X = x$, in fact it is *not* a probability measure, e.g., it can be $> 1$

- Can relate $f_X(x)$ to a probability using mean value theorem for integrals: Fix $x$ and some $\Delta x > 0$. Then provided $f_X$ is continuous over $(x, x + \Delta x]$,

  $$P\{X \in (x, x + \Delta x]\} = \int_x^{x+\Delta x} f_X(\alpha) \, d\alpha$$

  $$= f_X(c) \, \Delta x \text{ for some } x \leq c \leq x + \Delta x$$

  Now, if $\Delta x$ is sufficiently small, then

  $$P\{X \in (x, x + \Delta x]\} \approx f_X(x) \Delta x$$

- Notation: $X \sim f_X(x)$ means that $X$ has pdf $f_X(x)$
Famous Probability Density Functions

- **Uniform**: \( X \sim U[a, b] \) for \( b > a \) has the pdf

\[
f(x) = \begin{cases} 
\frac{1}{b-a} & \text{for } a \leq x \leq b \\
0 & \text{otherwise}
\end{cases}
\]

Uniform r.v. is commonly used to model quantization noise and finite precision computation error (roundoff error).

- **Exponential**: \( X \sim \text{Exp}(\lambda) \) for \( \lambda > 0 \) has the pdf

\[
f(x) = \begin{cases} 
\lambda e^{-\lambda x} & \text{for } x \geq 0 \\
0 & \text{otherwise}
\end{cases}
\]

Exponential r.v. is commonly used to model interarrival time in a queue, i.e., time between two consecutive packet or customer arrivals, service time in a queue, and lifetime of a particle, etc.
Example: Let $X \sim \text{Exp}(0.1)$ be the customer service time at a bank (in minutes). The person ahead of you has been served for 5 minutes. What is the probability that you will wait another 5 minutes or more before getting served?

We want to find $P\{X > 10 \mid X > 5\}$

Solution: By definition of conditional probability

$$P\{X > 10 \mid X > 5\} = \frac{P\{X > 10, X > 5\}}{P\{X > 5\}}$$

$$= \frac{P\{X > 10\}}{P\{X > 5\}}$$

$$= \frac{\int_{10}^{\infty} 0.1e^{-0.1x} \, dx}{\int_{5}^{\infty} 0.1e^{-0.1x} \, dx}$$

$$= \frac{e^{-1}}{e^{-0.5}} = e^{-0.5},$$

but $P\{X > 5\} = e^{-0.5}$, i.e., the conditional probability of waiting more than 5 additional minutes given that you have already waited more than 5 minutes is the same as the unconditional probability of waiting more than 5 minutes!

This is because the exponential r.v. is memoryless, which in general means that for any $0 \leq x_1 < x$,

$$P\{X > x \mid X > x_1\} = P\{X > x - x_1\}$$

To show this, consider

$$P\{X > x \mid X > x_1\} = \frac{P\{X > x, X > x_1\}}{P\{X > x_1\}}$$

$$= \frac{P\{X > x\}}{P\{X > x_1\}}$$

$$= \frac{\int_{x}^{\infty} \lambda e^{-\lambda x} \, dx}{\int_{x_1}^{\infty} \lambda e^{-\lambda x} \, dx}$$

$$= \frac{e^{-\lambda x}}{e^{-\lambda x_1}} = e^{-\lambda(x-x_1)}$$

$$= P\{X > x - x_1\}$$
• **Gaussian**: $X \sim \mathcal{N}(\mu, \sigma^2)$ has pdf

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

where $\mu$ is the *mean* and $\sigma^2$ is the *variance*

Gaussian r.v.s are frequently encountered in nature, e.g., thermal and shot noise in electronic devices are Gaussian, and very frequently used in modelling various social, biological, and other phenomena.

---

**Mean and Variance**

- A discrete (continuous) r.v. is completely specified by its pmf (pdf)
- It is often desirable to summarize the r.v. or predict its outcome in terms of one or a few numbers. What do we expect the value of the r.v. to be? What range of values around the mean do we expect the r.v. to take? Such information can be provided by the *mean* and *standard deviation* of the r.v.
- First we consider discrete r.v.s
- Let $X \sim p_X(x)$. The expected value (or mean) of $X$ is defined as

$$\mathbb{E}(X) = \sum_{x \in \mathcal{X}} xp_X(x)$$

Interpretations: If we view probabilities as relative frequencies, the mean would be the *weighted sum* of the relative frequencies. If we view probabilities as point masses, the mean would be the *center of mass* of the set of mass points.
Example: If the weather is good, which happens with probability 0.6, Alice walks the 2 miles to class at a speed \( \frac{5}{30} \) miles/hr, otherwise she rides a bus at speed 30 miles/hr. What is the expected time to get to class?

Solution: Define the discrete r.v. \( T \) to take the value \( 2\times\frac{1}{5} \) hr with probability 0.6 and \( 2\times\frac{1}{30} \) hr with probability 0.4. The expected value of \( T \)

\[
E(T) = 2/5 \times 0.6 + 2/30 \times 0.4 = 4/15 \text{ hr}
\]

• The **second moment** (or mean square or average power) of \( X \) is defined as

\[
E(X^2) = \sum_{x \in X} x^2 p_X(x)
\]

For the previous example, the second moment is

\[
E(T^2) = (2/5)^2 \times 0.6 + (2/30)^2 \times 0.4 = 22/225 \text{ hr}^2
\]

• The **variance** of \( X \) is defined as

\[
\text{Var}(X) = E [(X - E(X))^2] = \sum_{x \in X} (x - E(X))^2 p_X(x)
\]

Note that the \( \text{Var} \geq 0 \). The variance has the interpretation of the **moment of inertia** about the center of mass for a set of mass points

• The **standard deviation** of \( X \) is defined as \( \sigma_X = \sqrt{\text{Var}(X)} \)

• Variance in terms of mean and second moment: Expanding the square and using the linearity of sum, we obtain

\[
\text{Var}(X) = E [(X - E(X))^2] = \sum_{x} (x - E(X))^2 p_X(x)
\]

\[
= \sum_{x} (x^2 - 2x E(X) + [E(X)]^2) p_X(x)
\]

\[
= \sum_{x} x^2 p_X(x) - 2 E(X) \sum_{x} x p_X(x) + [E(X)]^2 \sum_{x} p_X(x)
\]

\[
= E(X^2) - 2[E(X)]^2 + [E(X)]^2
\]

\[
= E(X^2) - [E(X)]^2
\]

Note that since for any r.v., \( \text{Var}(X) \geq 0 \), \( E(X^2) \geq (E(X))^2 \)

So, for our example, \( \text{Var}(T) = 22/225 - (4/15)^2 = 0.02667 \).
Mean and Variance of Famous Discrete RVs

• Bernoulli r.v. \( X \sim \text{Bern}(p) \): The mean is

\[
E(X) = p \times 1 + (1 - p) \times 0 = p
\]

The second moment is

\[
E(X^2) = p \times 1^2 + (1 - p) \times 0^2 = p
\]

Thus the variance is

\[
\text{Var}(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1 - p)
\]

• Binomial r.v. \( X \sim \text{B}(n, p) \): It is not easy to find it using the definition. Later we use a much simpler method to show that

\[
E(X) = np
\]

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

Just \( n \) times the mean (variance) of a Bernoulli!

• Geometric r.v. \( X \sim \text{Geom}(p) \): The mean is

\[
E(X) = \sum_{k=1}^{\infty} kp(1 - p)^{k-1}
\]

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

\[
= \frac{p}{1 - p} \sum_{k=0}^{\infty} k(1 - p)^{k}
\]

\[
= \frac{p}{1 - p} \times \frac{1 - p}{p^2} = \frac{1}{p}
\]

Note: We found the series sum using the following trick:

Recall that for \(|a| < 1\), the sum of the geometric series

\[
\sum_{k=0}^{\infty} a^k = \frac{1}{1 - a}
\]
Suppose we want to evaluate
\[ \sum_{k=0}^{\infty} ka^k \]

Multiply both sides of the geometric series sum by \(a\) to obtain
\[ \sum_{k=0}^{\infty} a^{k+1} = \frac{a}{1-a} \]

Now differentiate both sides with respect to \(a\). The LHS gives
\[ \frac{d}{da} \sum_{k=0}^{\infty} a^{k+1} = \sum_{k=0}^{\infty} (k+1)a^k = \frac{1}{a} \sum_{k=0}^{\infty} ka^k \]

The RHS gives
\[ \frac{d}{da} \frac{a}{1-a} = \frac{1}{(1-a)^2} \]

Thus we have
\[ \sum_{k=0}^{\infty} ka^k = \frac{a}{(1-a)^2} \]

The second moment can be similarly evaluated to obtain
\[ \mathbb{E}(X^2) = \frac{2-p}{p^2} \]

The variance is thus
\[ \text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{1-p}{p^2} \]
- Poisson r.v. \( X \sim \text{Poisson}(\lambda) \): The mean is given by

\[
E(X) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda}
\]

\[
= e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{k!}
\]

\[
= e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!}
\]

\[
= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}
\]

\[
= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}
\]

\[
= \lambda
\]

Can show that the variance is also equal to \( \lambda \)

---

**Mean and Variance for Continuous RVs**

- Now consider a continuous r.v. \( X \sim f_X(x) \), the expected value (or mean) of \( X \) is defined as

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

This has the interpretation of the center of mass for a mass density

- The *second moment* and variance are similarly defined as:

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

\[
\text{Var}(X) = E\left[ (X - E(X))^2 \right] = E(X^2) - (E(X))^2
\]
• Uniform r.v. $X \sim U[a, b]$: The mean, second moment, and variance are given by

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

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

$$= \int_{a}^{b} x^2 \times \frac{1}{b - a} \, dx$$

$$= \frac{1}{b - a} \int_{a}^{b} x^2 \, dx$$

$$= \frac{1}{b - a} \times \frac{b^3 - a^3}{3} = \frac{b^3 - a^3}{3(b - a)}$$

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

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

Thus, for $X \sim U[0, 1]$, $E(X) = 1/2$ and $\text{Var} = 1/12$

• Exponential r.v. $X \sim \text{Exp}(\lambda)$: The mean and variance are given by

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

$$= \int_{0}^{\infty} x \lambda e^{-\lambda x} \, dx$$

$$= \left([-xe^{-\lambda x}]_{0}^{\infty} + \int_{0}^{\infty} e^{-\lambda x} \, dx \right) \text{ (integration by parts)}$$

$$= 0 - \frac{1}{\lambda} e^{-\lambda x}\bigg|_{0}^{\infty} = \frac{1}{\lambda}$$

$$E(X^2) = \int_{0}^{\infty} x^2 \lambda e^{-\lambda x} \, dx = \frac{2}{\lambda^2}$$

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

• For a Gaussian r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$, the mean is $\mu$ and the variance is $\sigma^2$ (will show this later using transforms)
Mean and Variance for Famous r.v.s

<table>
<thead>
<tr>
<th>Random Variable</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bern(p)</td>
<td>p</td>
<td>p(1 − p)</td>
</tr>
<tr>
<td>Geom(p)</td>
<td>1/p</td>
<td>(1 − p)/p²</td>
</tr>
<tr>
<td>B(n, p)</td>
<td>np</td>
<td>np(1 − p)</td>
</tr>
<tr>
<td>Poisson(λ)</td>
<td>λ</td>
<td>λ</td>
</tr>
<tr>
<td>U[a, b]</td>
<td>(a + b)/2</td>
<td>(b − a)²/12</td>
</tr>
<tr>
<td>Exp(λ)</td>
<td>1/λ</td>
<td>1/λ²</td>
</tr>
<tr>
<td>N(μ, σ²)</td>
<td>μ</td>
<td>σ²</td>
</tr>
</tbody>
</table>

Cumulative Distribution Function (cdf)

- For discrete r.v.s we use pmf’s, for continuous r.v.s we use pdf’s
- Many real-world r.v.s are mixed, that is, have both discrete and continuous components

Example: A packet arrives at a router in a communication network. If the input buffer is empty (happens with probability \( p \)), the packet is serviced immediately. Otherwise the packet must wait for a random amount of time as characterized by a pdf

Define the r.v. \( X \) to be the packet service time. \( X \) is neither discrete nor continuous

- There is a third probability function that characterizes all random variable types — discrete, continuous, and mixed. The cumulative distribution function or cdf \( F_X(x) \) of a random variable is defined by

\[
F_X(x) = P\{X \leq x\} \text{ for } x \in (-\infty, \infty)
\]
• Properties of the cdf:
  ○ Like the pmf (but unlike the pdf), the cdf is the probability of something. Hence, \(0 \leq F_X(x) \leq 1\)
  ○ Normalization axiom implies that
    \[
    F_X(\infty) = 1, \text{ and } F_X(-\infty) = 0
    \]
  ○ \(F_X(x)\) is monotonically nondecreasing, i.e., if \(a > b\) then \(F_X(a) \geq F_X(b)\)
  ○ The probability of any event can be easily computed from a cdf, e.g., for an interval \((a, b]\),
    \[
    P\{X \in (a, b]\} = P\{a < X \leq b\}
    = P\{X \leq b\} - P\{X \leq a\} \text{ (additivity)}
    = F_X(b) - F_X(a)
    \]
  ○ The probability of any outcome \(a\) is: \(P\{X = a\} = F_X(a) - F_X(a^-)\)

○ If a r.v. is discrete, its cdf consists of a set of steps

○ If \(X\) is a continuous r.v. with pdf \(f_X(x)\), then
    \[
    F_X(x) = P\{X \leq x\} = \int_{-\infty}^{x} f_X(\alpha) \, d\alpha
    \]
    So, the cdf of a continuous r.v. \(X\) is continuous
In fact the precise way to define a continuous random variable is through the continuity of its cdf. Further, if $F_X(x)$ is differentiable (almost everywhere), then

$$f_X(x) = \frac{dF_X(x)}{dx} = \lim_{\Delta x \to 0} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x}$$

$$= \lim_{\Delta x \to 0} \frac{P\{x < X \leq x + \Delta x\}}{\Delta x}$$

- The cdf of a mixed r.v. has the general form

$$\begin{align*}
F_X(x) &= \begin{cases} 
0 & \text{if } x < a \\
\int_a^x \frac{1}{b-a} \, d\alpha = \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\
1 & \text{if } x \geq b
\end{cases}
\end{align*}$$

**Examples**

- cdf of a uniform r.v.:

$$\begin{align*}
F_X(x) &= \begin{cases} 
0 & \text{if } x < a \\
\int_a^x \frac{1}{b-a} \, d\alpha = \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\
1 & \text{if } x \geq b
\end{cases}
\end{align*}$$
• cdf of an exponential r.v.:

\[ F_X(x) = 0, \quad X < 0, \text{ and } F_X(x) = 1 - e^{-\lambda x}, \quad x \geq 0 \]

![CDF and PDF of an exponential distribution](image)

- cdf for a mixed r.v.: Let \( X \) be the service time of a packet at a router. If the buffer is empty (happens with probability \( p \)), the packet is serviced immediately. If it is not empty, service time is described by an exponential pdf with parameter \( \lambda > 0 \).

The cdf of \( X \) is

\[ F_X(x) = \begin{cases} 
0 & \text{if } x < 0 \\
p & \text{if } x = 0 \\
p + (1 - p)(1 - e^{-\lambda x}) & \text{if } x > 0
\end{cases} \]
• cdf of a Gaussian r.v.: There is no nice closed form for the cdf of a Gaussian r.v., but there are many published tables for the cdf of a standard normal pdf \( \mathcal{N}(0,1) \), the \( \Phi \) function:

\[
\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}} d\xi
\]

More commonly, the tables are for the \( Q(x) = 1 - \Phi(x) \) function

Or, for the complementary error function: \( \text{erfc}(x) = 2Q(\sqrt{2}x) \) for \( x > 0 \)

As we shall see, the \( Q(\cdot) \) function can be used to quickly compute \( P\{X > a\} \) for any Gaussian r.v. \( X \)

**Functions of a Random Variable**

• We are often given a r.v. \( X \) with a known distribution (pmf, cdf, or pdf), and some function \( y = g(x) \) of \( x \), e.g., \( X^2, |X|, \cos X \), etc., and wish to specify \( Y \), i.e., find its pmf, if it is discrete, pdf, if continuous, or cdf

• Each of these functions is a random variable defined over the original experiment as \( Y(\omega) = g(X(\omega)) \). However, since we do not assume knowledge of the sample space or the probability measure, we need to specify \( Y \) directly from the pmf, pdf, or cdf of \( X \)

• First assume that \( X \sim p_X(x) \), i.e., a discrete random variable, then \( Y \) is also discrete and can be described by a pmf \( p_Y(y) \). To find it we find the probability of the inverse image \( \{\omega : Y(\omega) = y\} \) for every \( y \). Assuming \( \Omega \) is discrete:

\[
g(x_i) = y
\]
\[ p_Y(y) = P\{ \omega : Y(\omega) = y \} \]
\[ = \sum_{\{ \omega : g(X(\omega)) = y \}} P\{ \omega \} \]
\[ = \sum_{\{ x : g(x) = y \}} \sum_{\{ \omega : X(\omega) = x \}} P\{ \omega \} = \sum_{\{ x : g(x) = y \}} p_X(x) \]

Thus
\[ p_Y(y) = \sum_{\{ x : g(x) = y \}} p_X(x) \]

We can derive \( p_Y(y) \) directly from \( p_X(x) \) without going back to the original random experiment.

---

**Example:** Let the r.v. \( X \) be the maximum of two independent rolls of a four-sided die. Define a new random variable \( Y = g(X) \), where

\[ g(x) = \begin{cases} 
1 & \text{if } x \geq 3 \\
0 & \text{otherwise}
\end{cases} \]

Find the pmf for \( Y \)

**Solution:**

\[ p_Y(y) = \sum_{\{ x : g(x) = y \}} p_X(x) \]
\[ p_Y(1) = \sum_{\{ x : x \geq 3 \}} p_X(x) \]
\[ = \frac{5}{16} + \frac{7}{16} = \frac{3}{4} \]
\[ p_Y(0) = 1 - p_Y(1) = \frac{1}{4} \]
**Derived Densities**

- Let $X$ be a continuous r.v. with pdf $f_X(x)$ and $Y = g(X)$ such that $Y$ is a continuous r.v. We wish to find $f_Y(y)$.

- Recall derived pmf approach: Given $p_X(x)$ and a function $Y = g(X)$, the pmf of $Y$ is given by
  
  $$ p_Y(y) = \sum_{\{x: g(x) = y\}} p_X(x), $$

  i.e., $p_Y(y)$ is the sum of $p_X(x)$ over all $x$ that yield $g(x) = y$.

- Idea does not extend immediately to deriving pdfs, since pdfs are not probabilities, and we cannot add probabilities of points.
  
  But basic idea does extend to cdfs.

- Can first calculate the cdf of $Y$ as
  
  $$ F_Y(y) = P\{g(X) \leq y\} = \int_{\{x: g(x) \leq y\}} f_X(x) \, dx $$

  ![Diagram](diagram.png)

  \{x: g(x) \leq y\}

  - Then differentiate to obtain the pdf
    
    $$ f_Y(y) = \frac{dF_Y(y)}{dy} $$

    The hard part is typically getting the limits on the integral correct. Often they are obvious, but sometimes they are more subtle.
Example: Linear Functions

- Let $X \sim f_X(x)$ and $Y = aX + b$ for some $a > 0$ and $b$. (The case $a < 0$ is left as an exercise)

- To find the pdf of $Y$, we use the above procedure

\[\begin{align*}
F_Y(y) &= P\{Y \leq y\} = P\{aX + b \leq y\} \\
&= P\left\{X \leq \frac{y - b}{a}\right\} = F_X\left(\frac{y - b}{a}\right)
\end{align*}\]

Thus

\[f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{1}{a}f_X\left(\frac{y - b}{a}\right)\]

- Can show that for general $a \neq 0$,

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

- Example: $X \sim \text{Exp}(\lambda)$, i.e.,

\[f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0\]

Then

\[f_Y(y) = \begin{cases} 
\frac{\lambda}{a}e^{-\lambda(y-b)/a} & \text{if } \frac{y-b}{a} \geq 0 \\
0 & \text{otherwise}
\end{cases}\]
• Example: $X \sim \mathcal{N}(\mu, \sigma^2)$, i.e.,

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

Again setting $Y = aX + b$,

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

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

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

$$= \frac{1}{\sqrt{2\pi \sigma_Y^2}} e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}} \text{ for } -\infty < y < \infty$$

Therefore, $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

This result can be used to compute probabilities for an arbitrary Gaussian r.v. from knowledge of the distribution a $\mathcal{N}(0,1)$ r.v.

Suppose $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and we wish to find $P\{Y > y\}$

From the above result, we can express $Y = \sigma_Y X + \mu_Y$, where $X \sim \mathcal{N}(0,1)$

Thus

$$P\{Y > y\} = P\{\sigma_Y X + \mu_Y > y\}$$

$$= P\left\{ X > \frac{y - \mu_Y}{\sigma_Y} \right\}$$

$$= Q\left( \frac{y - \mu_Y}{\sigma_Y} \right)$$
Example: A Nonlinear Function

- John is driving a distance of 180 miles at constant speed that is uniformly distributed between 30 and 60 miles/hr. What is the pdf of the duration of the trip?

- Solution: Let $X$ be John’s speed, then

$$f_X(x) = \begin{cases} 
0 & \text{if } x < 30 \\
1/30 & \text{if } 30 \leq x \leq 60 \\
0 & \text{otherwise,}
\end{cases}$$

The duration of the trip is $Y = 180/X$

To find $f_Y(y)$ we first find $F_Y(y)$. Note that $\{y : Y \leq y\} = \{x : X \geq 180/y\}$, thus

$$F_Y(y) = \int_{180/y}^\infty f_X(x) \, dx$$

$$= \begin{cases} 
0 & \text{if } y \leq 3 \\
\int_{180/y}^{60} f_X(x) \, dx & \text{if } 3 < y \leq 6 \\
1 & \text{if } y > 6
\end{cases}$$

$$= \begin{cases} 
0 & \text{if } y \leq 3 \\
(2 - \frac{6}{y}) & \text{if } 3 < y \leq 6 \\
1 & \text{if } y > 6
\end{cases}$$

Differentiating, we obtain

$$f_Y(y) = \begin{cases} 
6/y^2 & \text{if } 3 \leq y \leq 6 \\
0 & \text{otherwise}
\end{cases}$$
• Let $g(x)$ be a monotonically increasing and differentiable function over its range. Then $g$ is invertible, i.e., there exists a function $h$, such that

$$y = g(x) \text{ if and only if } x = h(y)$$

Often this is written as $g^{-1}$. If a function $g$ has these properties, then $g(x) \leq y$ iff $x \leq h(y)$, and we can write

$$F_Y(y) = P\{g(X) \leq y\} = P\{X \leq h(y)\} = F_X(h(y)) = \int_{-\infty}^{h(y)} f_X(x) \, dx$$

Thus

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

• Generalizing the result to both monotonically increasing and decreasing functions yields

$$f_Y(y) = f_X(h(y)) \left| \frac{dh}{dy}(y) \right|$$

• Example: Recall the $X \sim U[30, 60]$ example with $Y = g(X) = 180/X$

The inverse is $X = h(Y) = 180/Y$

Applying the above formula in region of interest $Y \in [3, 6]$ (it is 0 outside) yields

$$f_Y(y) = f_X(h(y)) \left| \frac{dh}{dy}(y) \right| = \frac{1}{30} \frac{180}{y^2} = \frac{6}{y^2} \text{ for } 3 \leq y \leq 6$$
Another Example:

Suppose that $X$ has a pdf that is nonzero only in $[0, 1]$ and define $Y = g(X) = X^2$

In the region of interest the function is invertible and $X = \sqrt{Y}$

Applying the pdf formula, we obtain

$$f_Y(y) = f_X(h(y)) \left| \frac{dh}{dy}(y) \right|$$

$$= f_X(\sqrt{y}) \frac{1}{2\sqrt{y}}, \text{ for } 0 < y \leq 1$$

Personally I prefer the more fundamental approach, since I often forget this formula or mess up the signs
Two Discrete Random Variables – Joint PMFs

- As we have seen, one can define several r.v.s on the sample space of a random experiment. How do we jointly specify multiple r.v.s, i.e., be able to determine the probability of any event involving multiple r.v.s?

- We first consider two discrete r.v.s

- Let $X$ and $Y$ be two discrete random variables defined on the same experiment. They are completely specified by their **joint pmf**

\[
p_{X,Y}(x,y) = P\{X = x, Y = y\} \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y}
\]

- Clearly, $p_{X,Y}(x,y) \geq 0$, and $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y) = 1$

- Notation: We use $(X,Y) \sim p_{X,Y}(x,y)$ to mean that the two discrete r.v.s have the specified joint pmf
• The joint pmf can be described by a table

Example: Consider $X, Y$ with the following joint pmf $p_{X,Y}(x,y)$

<table>
<thead>
<tr>
<th></th>
<th>$X$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$1$</td>
<td>$1/16$</td>
<td>$0$</td>
<td>$1/8$</td>
<td>$1/16$</td>
</tr>
<tr>
<td>$Y$</td>
<td>$2$</td>
<td>$1/32$</td>
<td>$1/32$</td>
<td>$1/4$</td>
</tr>
<tr>
<td>$3$</td>
<td>$0$</td>
<td>$1/8$</td>
<td>$1/16$</td>
<td>$1/16$</td>
</tr>
<tr>
<td>$4$</td>
<td>$1/16$</td>
<td>$1/32$</td>
<td>$1/16$</td>
<td>$1/32$</td>
</tr>
</tbody>
</table>

Marginal PMFs

• Consider two discrete r.v.s $X$ and $Y$. They are described by their joint pmf $p_{X,Y}(x,y)$. We can also define their marginal pmfs $p_X(x)$ and $p_Y(y)$. How are these related?

• To find the marginal pmf of $X$, we use the law of total probability

$$p_X(x) = \sum_{y \in \mathcal{Y}} p(x, y) \text{ for } x \in \mathcal{X}$$

Similarly to find the marginal pmf of $Y$, we sum over $x \in \mathcal{X}$

• Example: Find the marginal pmfs for the previous example

<table>
<thead>
<tr>
<th></th>
<th>$X$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$1$</td>
<td>$1/16$</td>
<td>$0$</td>
<td>$1/8$</td>
<td>$1/16$</td>
</tr>
<tr>
<td>$Y$</td>
<td>$2$</td>
<td>$1/32$</td>
<td>$1/32$</td>
<td>$1/4$</td>
</tr>
<tr>
<td>$3$</td>
<td>$0$</td>
<td>$1/8$</td>
<td>$1/16$</td>
<td>$1/16$</td>
</tr>
<tr>
<td>$4$</td>
<td>$1/16$</td>
<td>$1/32$</td>
<td>$1/16$</td>
<td>$1/32$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$p_X(x)$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$1/16$</td>
<td>$1/32$</td>
<td>$1/16$</td>
</tr>
</tbody>
</table>
Conditional PMFs

- The conditional pmf of $X$ given $Y = y$ is defined as

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \text{ for } p_Y(y) \neq 0 \text{ and } x \in \mathcal{X}$$

Also, the conditional pmf of $Y$ given $X = x$ is

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \text{ for } p_X(x) \neq 0 \text{ and } y \in \mathcal{Y}$$

- For fixed $x$, $p_{Y|X}(y|x)$ is a pmf for $Y$

- Example: Find $p_{Y|X}(y|2)$ for the previous example

- Chain rule: Can write

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

- Bayes rule for pmfs: Given $p_X(x)$ and $p_{Y|X}(y|x)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we can find

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_{Y|X}(y|x)}{p_Y(y)}p_X(x) = \frac{p_{Y|X}(y|x)}{\sum_{x' \in \mathcal{X}} p_{X,Y}(x', y)}p_X(x), \text{ by total probability}$$

Using the chain rule, we obtain another version of Bayes rule

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)}{\sum_{x' \in \mathcal{X}} p_X(x')p_{Y|X}(y|x')}p_X(x)$$
Independence

- The random variables $X$ and $Y$ are said to be independent if for any events $A \in \mathcal{X}$ and $B \in \mathcal{Y}$

\[
P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}
\]

- Can show that the above definition is equivalent to saying that the r.v.s $X$ and $Y$ are independent if

\[
p_{X,Y}(x, y) = p_X(x)p_Y(y) \text{ for all } (x, y) \in \mathcal{X} \times \mathcal{Y}
\]

- Independence implies that $p_{X|Y}(x|y) = p_X(x)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$

Example: Binary Symmetric Channel

- Consider the following binary communication channel

\[
\begin{align*}
Z & \in \{0, 1\} \\
X & \in \{0, 1\} \\
Y & \in \{0, 1\}
\end{align*}
\]

The bit sent $X \sim \text{Bern}(p)$, the noise $Z \sim \text{Bern}(\epsilon)$, the bit received $Y = (X + Z) \mod 2 = X \oplus Z$, and $X$ and $Z$ are independent. Find

1. $p_{X|Y}(x|y)$,
2. $p_Y(y)$, and
3. the probability of error $P\{X \neq Y\}$
1. To find \( p_{X|Y}(x|y) \) we use Bayes rule

\[
p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)}{\sum_{x'} p_{Y|X}(y|x')p_X(x')}p_X(x)
\]

We know \( p_X(x) \). To find \( p_{Y|X} \), note that

\[
p_{Y|X}(y|x) = P\{Y = y|X = x\} = P\{X \oplus Z = y|X = x\} = P\{Z = y \oplus X|X = x\} = P\{Z = y \oplus x|X = x\} = P\{Z = y \oplus x\}, \text{ since } Z \text{ and } X \text{ are independent}
\]

So we have

\[
p_{Y|X}(0|0) = p_Z(0 \oplus 0) = p_Z(0) = 1 - \epsilon, \quad p_{Y|X}(0|1) = p_Z(0 \oplus 1) = p_Z(1) = \epsilon
\]

\[
p_{Y|X}(1|0) = p_Z(1 \oplus 0) = p_Z(1) = \epsilon, \quad p_{Y|X}(1|1) = p_Z(1 \oplus 1) = p_Z(0) = 1 - \epsilon
\]

Plugging in the Bayes rule equation, we obtain

\[
p_{X|Y}(0|0) = \frac{p_{Y|X}(0|0)}{p_{Y|X}(0|0)p_X(0) + p_{Y|X}(0|1)p_X(1)}p_X(0)
\]

\[
= \frac{(1 - \epsilon)(1 - p)}{(1 - \epsilon)(1 - p) + \epsilon p}
\]

\[
p_{X|Y}(1|0) = \frac{p_{Y|X}(0|1)}{p_{Y|X}(0|0)p_X(0) + p_{Y|X}(0|1)p_X(1)}p_X(1)
\]

\[
= \frac{\epsilon p}{(1 - \epsilon)(1 - p) + \epsilon p}
\]

\[
p_{X|Y}(0|1) = \frac{p_{Y|X}(1|0)}{p_{Y|X}(1|0)p_X(0) + p_{Y|X}(1|1)p_X(1)}p_X(0)
\]

\[
= \frac{\epsilon(1 - p)}{\epsilon(1 - p) + (1 - \epsilon)p}
\]

\[
p_{X|Y}(1|1) = \frac{p_{Y|X}(1|1)}{p_{Y|X}(1|0)p_X(0) + p_{Y|X}(1|1)p_X(1)}p_X(1)
\]

\[
= \frac{(1 - \epsilon)p}{\epsilon(1 - p) + (1 - \epsilon)p}
\]
2. We already found $p_Y(y)$:  
\[ p_Y(1) = \epsilon(1 - p) + (1 - \epsilon)p \]

3. Now to find the probability of error $P\{X \neq Y\}$, consider  
\[ P\{X \neq Y\} = p_{X,Y}(0,1) + p_{X,Y}(1,0) \]
\[ = p_{Y|X}(1|0)p_X(0) + p_{Y|X}(0|1)p_X(1) \]
\[ = \epsilon(1 - p) + \epsilon p \]
\[ = \epsilon \]

An interesting special case is when $\epsilon = 1/2$

Here, $P\{X \neq Y\} = 1/2$, which is the worst possible (no information is sent), and  
\[ p_Y(0) = \frac{1}{2}p + \frac{1}{2}(1 - p) = \frac{1}{2} = p_Y(1), \]

i.e., $Y \sim \text{Bern}(1/2)$, independent of the value of $p$.

Also in this case, the bit sent $X$ and the bit received $Y$ are independent (check this).

---

### Two Continuous Random Variables – Joint PDFs

- Two continuous r.v.s defined over the same experiment are *jointly continuous* if they take on a continuum of values each with probability 0. They are completely specified by a joint pdf $f_{X,Y}$ such that for any event $A \in (-\infty, \infty)^2$,  
\[ P\{(X, Y) \in A\} = \int_{(x,y) \in A} f_{X,Y}(x, y) \, dx \, dy \]

For example, for a rectangular area  
\[ P\{a < X \leq b, c < Y \leq d\} = \int_c^d \int_a^b f_{X,Y}(x, y) \, dx \, dy \]

- Properties of a joint pdf $f_{X,Y}$:
  1. $f_{X,Y}(x, y) \geq 0$
  2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx \, dy = 1$

- Again joint pdf is *not* a probability. Can relate it to probability as  
\[ P\{x < X \leq x + \Delta x, y < Y \leq y + \Delta y\} \approx f_{X,Y}(x, y) \, \Delta x \, \Delta y \]
Marginal PDF

- The *Marginal* pdf of $X$ can be obtained from the joint pdf by integrating the joint over the other variable $y$

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

  This follows by the law of total probability. To see this, consider

  \[
  f_X(x) = \lim_{\Delta x \to 0} \frac{\mathbb{P}\{x < X \leq x + \Delta x\}}{\Delta x} \\
  = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \lim_{\Delta y \to 0} \sum_{n=-\infty}^{\infty} \mathbb{P}\{x < X \leq x + \Delta x, n\Delta y < Y \leq (n+1)\Delta y\} \\
  = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy \, \Delta x = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy
  \]
Example

• Let \((X, Y) \sim f(x, y)\), where

\[
f(x, y) = \begin{cases} 
c & \text{if } x, y \geq 0, \text{ and } x + y \leq 1 \\
0 & \text{otherwise}
\end{cases}
\]

1. Find \(c\)
2. Find \(f_Y(y)\)
3. Find \(P\{X \geq \frac{1}{2}Y\}\)

• Solution:

1. To find \(c\), note that

\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = 1,
\]

thus \(\frac{1}{2}c = 1\), or \(c = 2\)

2. To find \(f_Y(y)\), we use the law of total probability

\[
f_Y(y) = \int_{-\infty}^{\infty} f(x, y) \, dx
\]

\[
= \begin{cases} 
\int_{0}^{1-y} 2 \, dx & 0 \leq y \leq 1 \\
0 & \text{otherwise}
\end{cases}
\]

\[
= \begin{cases} 
2(1 - y) & 0 \leq y \leq 1 \\
0 & \text{otherwise}
\end{cases}
\]
3. To find the probability of the set \( \{ X \geq \frac{1}{2} Y \} \) we first sketch it

\[
P \left\{ X \geq \frac{1}{2} Y \right\} = \int_{ \{(x,y): x \geq \frac{1}{2} y \} } f_{X,Y}(x,y) \, dx \, dy
\]

from the figure we find that

\[
P \left\{ X \geq \frac{1}{2} Y \right\} = \int_{\frac{2}{3}}^{1} \int_{\frac{y}{2}}^{1-y} 2 \, dx \, dy = \frac{2}{3}
\]

---

**Example: Buffon’s Needle Problem**

- The plane is ruled with equidistant parallel lines at distance \( d \) apart. Throw needle of length \( l < d \) at random. What is the probability that it will intersect one of the lines?

- Solution:

  Let \( X \) be the distance from the midpoint of the needle to the nearest of the parallel lines, and \( \Theta \) be the acute angle determined by a line through the needle and the nearest parallel line
\((X, \Theta)\) are uniformly distributed within the rectangle \([0, d/2] \times [0, \pi/2]\), thus

\[
f_{X,\Theta}(x, \theta) = \frac{4}{\pi d}, \quad x \in [0, d/2], \ \theta \in [0, \pi/2]
\]

The needle intersects a line iff \(X < \frac{l}{2} \sin \Theta\)

The probability of intersection is:

\[
P\left\{ X < \frac{l}{2} \sin \Theta \right\} = \int \int_{\{(x, \theta) : x < \frac{l}{2} \sin \theta \}} f_{X,\Theta}(x, \theta) \, dx \, d\theta
\]

\[
= \frac{4}{\pi d} \int_0^{\pi/2} \int_0^{\frac{l}{2} \sin \theta} dx \, d\theta
\]

\[
= \frac{4}{\pi d} \int_0^{\pi/2} \frac{l}{2} \sin \theta \, d\theta
\]

\[
= \frac{2l}{\pi d}
\]

---

**Example: Darts**

- Throw a dart on a disk of radius \(r\). Probability on the coordinates \((X, Y)\) is described by a uniform pdf on the disk:

\[
f_{X,Y}(x, y) = \begin{cases} \frac{1}{\pi r^2}, & \text{if } x^2 + y^2 \leq r^2 \\ 0, & \text{otherwise} \end{cases}
\]

Find the marginal pdfs

- Solution: To find the pdf of \(Y\) (same as \(X\), consider

\[
f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx = \frac{1}{\pi r^2} \int_{\{x : x^2 + y^2 \leq r^2\}} dx
\]

\[
= \frac{1}{\pi r^2} \int_{-\sqrt{r^2-y^2}}^{\sqrt{r^2-y^2}} dx = \begin{cases} \frac{2}{\pi r^2} \sqrt{r^2 - y^2}, & \text{if } |y| \leq r \\ 0, & \text{otherwise} \end{cases}
\]
Conditional pdf

- Let $X$ and $Y$ be continuous random variables with joint pdf $f_{X,Y}(x,y)$, we define the conditional pdf of $Y$ given $X$ as
  \[ f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \text{ for } f_X(x) \neq 0 \]

- Note that, for a fixed $X = x$, $f_{Y|X}(y|x)$ is a legitimate pdf on $Y$ – it is nonnegative and integrates to 1.

- We want the conditional pdf to be interpreted as:
  \[ f_{Y|X}(y|x)\Delta y \approx P\{y < Y \leq y + \Delta y | X = x\} \]

  The RHS can be interpreted as a limit
  \[
P\{y < Y \leq y + \Delta y | X = x\} = \lim_{\Delta x \to 0} \frac{P\{y < Y \leq y + \Delta y, x < X \leq x + \Delta x\}}{P\{x < X \leq x + \Delta x\}} \\
  \approx \lim_{\Delta x \to 0} \frac{f_{X,Y}(x,y) \Delta y}{f_X(x) \Delta x} = \frac{f_{X,Y}(x,y)}{f_X(x)} \Delta y
  \]

Example: Let
  \[ f(x,y) = \begin{cases} 
  2, & x,y \geq 0, x+y \leq 1 \\
  0, & \text{otherwise}
  \end{cases} \]

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

Solution: We already know that $f_Y(y) = \begin{cases} 
  2(1-y), & 0 \leq y \leq 1 \\
  0, & \text{otherwise}
  \end{cases}$

Therefore
  \[ f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} 
  \frac{1}{1-y}, & x,y \geq 0, x+y \leq 1, y < 1 \\
  0, & \text{otherwise}
  \end{cases} \]
• Chain rule: \( p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x) \)


\[ f_{X,Y}(x,y) = f_X(x)f_Y(y) \text{ for all } x, y \]

It can be shown that this definition is equivalent to saying that \( X \) and \( Y \) are independent if for any two events \( A, B \subset (-\infty, \infty) \)

\[ P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\} \]

• Example: Are \( X \) and \( Y \) in the previous example independent?

• Bayes rule for densities: Given \( f_X(x) \) and \( f_{Y|X}(y|x) \), we can find

\[
f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)}{f_Y(y)}f_X(x) \\
= \frac{f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_{X,Y}(u,y) du}f_X(x) \\
= \frac{f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(u)f_{Y|X}(y|u) du}f_X(x)
\]
**Example:** Let $\Lambda \sim U[0, 1]$, and the conditional pdf of $X$ given $\Lambda = \lambda$

$$f_{X|\Lambda}(x|\lambda) = \lambda e^{-\lambda x}, \quad 0 \leq \lambda \leq 1,$$

i.e., $X|\{\Lambda = \lambda\} \sim \text{Exp}(\lambda)$. Now, given that $X = 3$, find $f_{\Lambda|X}(\lambda|3)$

**Solution:** We use Bayes rule

$$f_{\Lambda|X}(\lambda|3) = \frac{f_{X|\Lambda}(3|\lambda)f_{\Lambda}(\lambda)}{\int_0^1 f_{\Lambda}(u)f_{X|\Lambda}(3|u) \, du}$$

$$= \begin{cases} \frac{\lambda e^{-3\lambda}}{\frac{1}{1 - 4e^{-3}}}, & 0 \leq \lambda \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

---

**Joint cdf**

- If $X$ and $Y$ are two r.v.s over the same experiment, they are completely specified by their **joint cdf**

$$F_{X,Y}(x, y) = P\{X \leq x, Y \leq y\} \text{ for } x, y \in (-\infty, \infty)$$
Properties of the joint cdf:

1. $F_{X,Y}(x, y) \geq 0$
2. $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$ whenever $x_1 \leq x_2$ and $y_1 \leq y_2$
3. $\lim_{x,y \to \infty} F_{X,Y}(x, y) = 1$
4. $\lim_{y \to \infty} F_{X,Y}(x, y) = F_X(x)$ and $\lim_{x \to \infty} F_{X,Y}(x, y) = F_Y(y)$
5. $\lim_{y \to -\infty} F_{X,Y}(x, y) = 0$ and $\lim_{x \to -\infty} F(x, y) = 0$
6. The probability of any set can be determined from the joint cdf, for example,

\[
P\{a < X \leq b, c < Y \leq d\} = F(b, d) - F(a, d) - F(b, c) + F(a, c)
\]

If $X$ and $Y$ are continuous random variables having a joint pdf $f_{X,Y}(x, y)$, then

\[
F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) \, du \, dv \text{ for } x, y \in (-\infty, \infty)
\]

Moreover, if $F_{X,Y}(x, y)$ is differentiable in both $x$ and $y$, then

\[
f_{X,Y}(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} = \lim_{\Delta x, \Delta y \to 0} \frac{P\{x < X \leq x + \Delta x, y < Y \leq y + \Delta y\}}{\Delta x \Delta y}
\]

Two random variables are independent if

\[
F_{X,Y}(x, y) = F_X(x) F_Y(y)
\]
Functions of Two Random Variables

• Let \( X \) and \( Y \) be two r.v.s with known pdf \( f_{X,Y}(x,y) \) and \( Z = g(x,y) \) be a function of \( X \) and \( Y \). We wish to find \( f_Z(z) \)

• We use the same procedure as before: First calculate the cdf of \( Z \), then differentiate it to find \( f_Z(z) \)

• Example: Max and Min of Independent Random Variables

Let \( X \sim f_X(x) \) and \( Y \sim f_Y(y) \) be independent, and define

\[
U = \max\{X, Y\}, \quad \text{and} \quad V = \min\{X, Y\}
\]

Find the pdfs of \( U \) and \( V \)

• Solution: To find the pdf of \( U \), we first find its cdf

\[
F_U(u) = P\{U \leq u\}
\]

\[
= P\{X \leq u, Y \leq u\}
\]

\[
= F_X(u)F_Y(u), \quad \text{by independence}
\]

Now, to find the pdf, we take the derivative w.r.t. \( u \)

\[
f_U(u) = f_X(u)F_Y(u) + f_Y(u)F_X(u)
\]

For example, if \( X \) and \( Y \) are uniformly distributed between 0 and 1,

\[
f_U(u) = 2u \quad \text{for} \quad 0 \leq u \leq 1
\]

Next, to find the pdf of \( V \), consider

\[
F_V(v) = P\{V \leq v\}
\]

\[
= 1 - P\{V > v\}
\]

\[
= 1 - P\{X > v, Y > v\}
\]

\[
= 1 - (1 - F_X(v))(1 - F_Y(v))
\]

\[
= F_X(v) + F_Y(v) - F_X(v)F_Y(v),
\]

thus

\[
f_V(v) = f_X(v) + f_Y(v) - f_X(v)f_Y(v) - f_Y(v)f_X(v)
\]

For example, if \( X \sim \text{Exp}(\lambda_1) \) and \( Y \sim \text{Exp}(\lambda_2) \), then \( V \sim \text{Exp}(\lambda_1 + \lambda_2) \)
• Let $X$ and $Y$ be independent r.v.s with known distributions. We wish to find the distribution of their sum $W = X + Y$

• First assume $X \sim p_X(x)$ and $Y \sim p_Y(y)$ are independent integer-valued r.v.s, then for any integer $w$, the pmf of their sum

$$p_W(w) = P\{X + Y = w\} = \sum_{\{(x,y): x+y=w\}} P\{X = x, Y = y\}$$

$$= \sum_x P\{X = x, Y = w-x\}$$

$$= \sum_x P\{X = x\}P\{Y = w-x\}, \text{ by independence}$$

$$= \sum_x p_X(x)p_Y(w-x)$$

This is the discrete convolution of the two pmfs

For example, let $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ be independent, then the pmf of their sum

$$p_W(w) = \sum_{x=-\infty}^{\infty} p_X(x)p_Y(w-x)$$

$$= \sum_{x=0}^{w} p_X(x)p_Y(w-x) \text{ for } w = 0, 1, \ldots$$

$$= \sum_{x=0}^{w} \frac{\lambda_1^x}{x!} e^{-\lambda_1} \frac{\lambda_2^{w-x}}{(w-x)!} e^{-\lambda_2}$$

$$= \frac{(\lambda_1 + \lambda_2)^w}{w!} e^{-(\lambda_1+\lambda_2)} \sum_{x=0}^{w} \binom{w}{x} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^x \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{w-x}$$

$$= \frac{(\lambda_1 + \lambda_2)^w}{w!} e^{-(\lambda_1+\lambda_2)}$$

Thus $W = X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$

In general a Poisson r.v. with parameter $\lambda$ can be written as the sum of any number of independent Poisson($\lambda_i$) r.v.s, so long as $\sum \lambda_i = \lambda$. This property of a distribution is called infinite divisibility.
Now, let’s assume that $X \sim f_X(x)$, and $Y \sim f_Y(y)$ are independent continuous r.v.s. We wish to find the pdf of their sum $W = X + Y$. To do so, first note that

$$P\{W \leq w | X = x\} = P\{X + Y \leq w | X = x\} = P\{x + Y \leq w | X = x\} = P\{x + Y \leq w\}, \text{ by independence}$$

Thus

$$f_{W|X}(w|x) = f_Y(w - x), \text{ a very useful result}$$

Now, to find the pdf of $W$, consider

$$f_W(w) = \int_{-\infty}^{\infty} f_{W,X}(w,x) \, dx$$

$$= \int_{-\infty}^{\infty} f_X(x)f_{W|X}(w|x) \, dx = \int_{-\infty}^{\infty} f_X(x)f_Y(w - x) \, dx$$

This is the convolution of $f_X(x)$ and $f_Y(y)$.

**Example:** Assume that $X \sim U[0, 1]$ and $Y \sim U[0, 1]$ are independent r.v.s. Find the pdf of their sum $W = X + Y$.  

**Solution:** To find the pdf of the sum, we convolve the two pdfs

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x)f_Y(w - x) \, dx$$

$$= \begin{cases} 
  w, & \text{if } 0 \leq w \leq 1 \\
  2 - w, & \text{if } 1 < w \leq 2 \\
  0, & \text{otherwise}
\end{cases}$$

**Example:** If $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then their sum $W \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$, i.e., Gaussian is also an infinitely divisible distribution—any Gaussian r.v. can be written as the sum of any number of independent Gaussians as long as their means sum to its mean and their variances sum to its variance (will prove this using transforms later).
One Discrete and One Continuous RVs

- Let Θ be a discrete random variable with pmf \( p_Θ(\theta) \).
- For each \( \Theta = \theta \), such that \( p_Θ(\theta) \neq 0 \), let \( Y \) be a continuous random variable with conditional pdf \( f_{Y|\Theta}(y|\theta) \).
- The conditional pmf of \( \Theta \) given \( Y \) can be defined as a limit
  \[
  p_Θ|Y(\theta|y) = \lim_{\Delta y \to 0} \frac{P\{\Theta = \theta, y < Y \leq y + \Delta y\}}{P\{y < Y \leq y + \Delta y\}}
  = \lim_{\Delta y \to 0} \frac{p_Θ(\theta)f_{Y|\Theta}(y|\theta)\Delta y}{f_Y(y)\Delta y}
  = \frac{f_{Y|\Theta}(y|\theta)}{f_Y(y)} p_Θ(\theta)
  \]
- So we obtain yet another version of Bayes rule
  \[
  p_Θ|Y(\theta|y) = \frac{f_{Y|\Theta}(y|\theta)}{\sum_{\theta'} p_Θ(\theta') f_{Y|\Theta}(y|\theta')} p_Θ(\theta)
  \]

Example: Additive Gaussian Noise Channel

- Consider the following communication channel model
  \[
  Z \sim \mathcal{N}(0, N)
  \]
  \[
  \Theta \quad \oplus \quad Y
  \]
  where the signal sent
  \[
  \Theta = \begin{cases} 
    +1, & \text{with probability } p \\
    -1, & \text{with probability } 1 - p,
  \end{cases}
  \]
  the signal received (also called observation) \( Y = \Theta + Z \), and \( \Theta \) and \( Z \) are independent.
  Given \( Y = y \) is received (observed), find the a posteriori pmf of \( \Theta \), \( p_Θ|Y(\theta|y) \).
• Solution: We use Bayes rule

\[
p_{\Theta|Y}(\theta|y) = \frac{f_{Y|\Theta}(y|\theta)p_{\Theta}(\theta)}{\sum_{\theta'} p_{\Theta}(\theta')f_{Y|\Theta}(y|\theta')}
\]

We know \(p_{\Theta}(\theta)\). To find \(f_{Y|\Theta}(y|\theta)\), consider

\[
P\{Y \leq y|\Theta = 1\} = P\{\Theta + Z \leq y|\Theta = 1\}
\]
\[
= P\{Z \leq y - \Theta|\Theta = 1\}
\]
\[
= P\{Z \leq y - 1|\Theta = 1\}
\]
\[
= P\{Z \leq y - 1\}, \text{ by independence of } \Theta \text{ and } Z
\]

Therefore, \(Y|\{\Theta = +1\} \sim \mathcal{N}(+1, N)\). Also, \(Y|\{\Theta = -1\} \sim \mathcal{N}(-1, N)\)

Thus

\[
p_{\Theta|Y}(1|y) = \frac{pe^y}{pe^y + (1-p)e^{-y}} \text{ for } -\infty < y < \infty
\]

• Now, let \(p = 1/2\). Suppose the receiver decides that the signal transmitted is 1 if \(Y > 0\), otherwise he decides that it is a \(-1\). What is the probability of decision error?

• Solution: First we plot the conditional pdfs

This decision rule make sense, since you decide that the signal transmitted is 1 if \(f_{Y|\Theta}(y|1) > f_{Y|\Theta}(y|-1)\)

An error occurs if

- \(\Theta = 1\) is transmitted and \(Y \leq 0\), or
- \(\Theta = -1\) is transmitted and \(Y > 0\)
But if \( \Theta = 1 \), then \( Y \leq 0 \) iff \( Z < -1 \), and if \( \Theta = -1 \), then \( Y > 0 \) iff \( Z > 1 \).

Thus the probability of error is

\[
P\{\text{error}\} = P\{\Theta = 1, Y \leq 0 \text{ or } \Theta = -1, Y > 0\}
= P\{\Theta = 1, Y \leq 0\} + P\{\Theta = -1, Y > 0\}
= P\{\Theta = 1\}P\{Y \leq 0|\Theta = 1\} + P\{\Theta = -1\}P\{Y > 0|\Theta = -1\}
= \frac{1}{2}P\{Z < -1\} + \frac{1}{2}P\{Z > 1\}
= Q\left(\frac{1}{\sqrt{N}}\right)
\]

**Summary: Total Probability and Bayes Rule**

- **Law of total probability:**
  - events: \( P(B) = \sum_i P(A_i \cap B) \), \( A_i \)’s partition \( \Omega \)
  - pmf: \( p_X(x) = \sum_y p(x, y) \)
  - pdf: \( f_X(x) = \int f_{X,Y}(x,y) \, dy \)
  - mixed: \( f_Y(y) = \sum_\theta p_\Theta(\theta)f_{Y|\Theta}(y|\theta) \), \( p_\Theta(\theta) = \int f_Y(y)p_{\Theta|Y}(\theta|y) \, dy \)

- **Bayes rule:**
  - events: \( P(A_j|B) = \frac{P(B|A_j)p(A_j)}{\sum_i P(B|A_i)p(A_i)} \), \( A_i \)’s partition \( \Omega \)
  - pmf: \( p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{x'} p_{Y|X}(y|x')p_X(x')} \)
  - pdf: \( f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)}{\int f_X(x')f_{Y|X}(y|x') \, dx'}f_X(x) \)
  - mixed: \( p_{\Theta|Y}(\theta|y) = \frac{f_{Y|\Theta}(\theta|y)p_{\Theta}(\theta)}{\int f_Y(y)p_{\Theta|Y}(\theta|y') \, dy'}f_Y(y) \)
More Than Two RVs

- Let $X_1, X_2, \ldots, X_n$ be random variables (defined over the same experiment).
- If the r.v.s are discrete then they can be jointly specified by their joint pmf
  \[ p_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n), \text{ for all } (x_1, x_2, \ldots, x_n) \in X_1 \times X_2 \times \ldots \times X_n \]
- If the r.v.s are jointly continuous, then they can be specified by the joint pdf
  \[ f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n), \text{ for all } (x_1, x_2, \ldots, x_n) \]
- Marginal pdf (or pmf) is the joint pdf (or pmf) for a subset of $\{X_1, \ldots, X_n\}$;
  e.g. for three r.v.s $X_1, X_2, X_3$, the marginals are $f_{X_i}(x_i)$ and $f_{X_i, X_j}(x_i, x_j)$ for $i \neq j$
- The marginals can be obtained from the joint in the usual way, e.g. for the $n = 3$ example
  \[ f_{X_1, X_2}(x_1, x_2) = \int_{-\infty}^{\infty} f_{X_1, X_2, X_3}(x_1, x_2, x_3) \, dx_3 \]
- Conditional pmf or pdf can be defined in the usual way, e.g. the conditional pdf
  of $(X_{k+1}, X_{k+2}, \ldots, X_n)$ given $(X_1, X_2, \ldots, X_k)$ is
  \[ f_{X_{k+1}, \ldots, X_n|X_1, \ldots, X_k}(x_{k+1}, \ldots, x_n|x_1, \ldots, x_k) = \frac{f_{X_1, \ldots, X_n}(x_1, \ldots, x_n)}{f_{X_1, \ldots, X_k}(x_1, \ldots, x_k)} \]
- Chain rule: We can write
  \[ f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_1, X_2}(x_3|x_1, x_2) \ldots \]
  \[ \ldots f_{X_n|X_1, \ldots, X_{n-1}}(x_n|x_1, \ldots, x_{n-1}) \]
- In general $X_1, X_2, \ldots, X_n$ are completely specified by their joint cdf
  \[ F_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = P\{X_1 \leq x_1, \ldots, X_n \leq x_n\}, \text{ for all } (x_1, \ldots, x_n) \]
Independence and Conditional Independence

- Independence is defined in the usual way: \( X_1, X_2, \ldots, X_n \) are said to be independent iff
  \[
  f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) \text{ for all } (x_1, x_2, \ldots, x_n)
  \]

- Important special case, \( i.i.d. \) r.v.s: \( X_1, X_2, \ldots, X_n \) are said to be independent and identically distributed (i.i.d.) if they are independent and have the same marginal, e.g. if we flip a coin \( n \) times independently we can generate \( X_1, X_2, \ldots, X_n \) i.i.d. \( \text{Bern}(p) \) r.v.s

- The r.v.s \( X_1 \) and \( X_3 \) are said to be conditionally independent given \( X_2 \) iff
  \[
  f_{X_1, X_3|X_2}(x_1, x_3|x_2) = f_{X_1|X_2}(x_1|x_2)f_{X_3|X_2}(x_3|x_2) \text{ for all } (x_1, x_2, x_3)
  \]

- Conditional independence does not necessarily imply or is implied by independence, i.e., \( X_1 \) and \( X_3 \) independent given \( X_2 \) does not necessarily mean that \( X_1 \) and \( X_3 \) are independent (or vice versa)

**Example: Series Binary Symmetric Channels:**

Here \( X_1 \sim \text{Bern}(p) \), \( Z_1 \sim \text{Bern}(\epsilon_1) \), and \( Z_2 \sim \text{Bern}(\epsilon_2) \) are independent, and \( X_3 = X_1 + Z_1 + Z_2 \mod 2 \)

- In general \( X_1 \) and \( X_3 \) are not independent
- \( X_1 \) and \( X_3 \) are conditionally independent given \( X_2 \)
- Also \( X_1 \) and \( Z_1 \) are independent, but not conditionally independent given \( X_2 \)

**Example Coin with Random Bias:** Consider a coin with random bias \( P \sim f_P(p) \). Flip it \( n \) times independently to generate the r.v.s \( X_1, X_2, \ldots, X_n \) (\( X_i = 1 \) if \( i\)-th flip is heads, 0 otherwise)

- The r.v.s \( X_1, X_2, \ldots, X_n \) are not independent
- However, \( X_1, X_2, \ldots, X_n \) are conditionally independent given \( P \) — in fact, for any \( P = p \), they are i.i.d. \( \text{Bern}(p) \)
Definition

- We already introduced the notion of expectation (mean) of a r.v.
- We generalize this definition and discuss it in more depth
- Let $X \in \mathcal{X}$ be a discrete r.v. with pmf $p_X(x)$ and $g(x)$ be a function of $x$. The expectation or expected value of $g(X)$ is defined as
  \[ E(g(X)) = \sum_{x \in \mathcal{X}} g(x)p_X(x) \]
- For a continuous r.v. $X \sim f_X(x)$, the expected value of $g(X)$ is defined as
  \[ E(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x) \, dx \]
- Examples:
  - $g(X) = c$, a constant, then $E(g(X)) = c$
  - $g(X) = X$, $E(X) = \sum_{x} xp_X(x)$ is the mean of $X$
  - $g(X) = X^k$, $E(X^k)$ is the $k$th moment of $X$
  - $g(X) = (X - E(X))^2$, $E[(X - E(X))^2]$ is the variance of $X$
Expectation is linear, i.e., for any constants $a$ and $b$

$$E[ag_1(X) + bg_2(X)] = aE(g_1(X)) + bE(g_2(X))$$

Examples:

- $E(aX + b) = aE(X) + b$
- $\text{Var}(aX + b) = a^2\text{Var}(X)$
  
  Proof: From the definition

$$\text{Var}(aX + b) = E[(aX + b - E(aX + b))^2]$$

$$= E[(aX + b - aE(X) - b)^2]$$

$$= E[a^2(X - E(X))^2]$$

$$= a^2E((X - E(X))^2)$$

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

### Fundamental Theorem of Expectation

- Theorem: Let $X \sim p_X(x)$ and $Y = g(X) \sim p_Y(y)$, then

$$E(Y) = \sum_{y \in Y} yp_Y(y) = \sum_{x \in X} g(x)p_X(x) = E(g(X))$$

- The same formula holds for $f_Y(y)$ using integrals instead of sums
- Conclusion: $E(Y)$ can be found using either $f_X(x)$ or $f_Y(y)$. It is often much easier to use $f_X(x)$ than to first find $f_Y(y)$ then find $E(Y)$
- Proof: We prove the theorem for discrete r.v.s. Consider

$$E(Y) = \sum_{y} yp_Y(y)$$

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

$$= \sum_{y} \sum_{\{x: g(x) = y\}} yp_X(x)$$

$$= \sum_{y} \sum_{\{x: g(x) = y\}} g(x)p_X(x) = \sum_{x} g(x)p_X(x)$$
• Let \((X, Y) \sim f_{X,Y}(x, y)\) and let \(g(x, y)\) be a function of \(x\) and \(y\). The expectation of \(g(X, Y)\) is defined as

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

The function \(g(X, Y)\) may be \(X, Y, X^2, X + Y\), etc.

• The correlation of \(X\) and \(Y\) is defined as \(E(XY)\)

• The covariance of \(X\) and \(Y\) is defined as

\[
\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY - XE(Y) - YE(X) + E(X)E(Y)) = E(XY) - E(X)E(Y)
\]

Note that if \(X = Y\), then \(\text{Cov}(X, Y) = \text{Var}(X)\)

• Example: Let

\[
f(x, y) = \begin{cases} 
2 & \text{for } x, y \geq 0, x + y \leq 1 \\
0 & \text{otherwise}
\end{cases}
\]

Find \(E(X), \text{Var}(X), \text{and Cov}(X, Y)\)

Solution: The mean is

\[
E(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x, y) \, dy \, dx
\]

\[
= \int_{0}^{1} \int_{0}^{1-x} 2x \, dy \, dx = 2 \int_{0}^{1} (1 - x) x \, dx = \frac{1}{3}
\]

To find the variance, we first find the second moment

\[
E(X^2) = \int_{0}^{1} \int_{0}^{1-x} 2x^2 \, dy \, dx = 2 \int_{0}^{1} (1 - x)x^2 \, dx = \frac{1}{6}
\]

Thus,

\[
\text{Var}(X) = E(X^2) - (E(X))^2 = \frac{1}{6} - \frac{1}{9} = \frac{1}{18}
\]
The covariance of $X$ and $Y$ is

$$\text{Cov}(X, Y) = 2 \int_0^1 \int_0^{1-x} xy \, dy \, dx - \text{E}(X) \text{E}(Y)$$

$$= \int_0^1 x(1-x)^2 \, dx - \frac{1}{9} = \frac{1}{12} - \frac{1}{9} = -\frac{1}{36}$$

---

**Independence and Uncorrelation**

- Let $X$ and $Y$ be independent r.v.s and $g(X)$ and $h(Y)$ be functions of $X$ and $Y$, respectively, then

$$\text{E}(g(X)h(Y)) = \text{E}(g(X)) \text{E}(h(Y))$$

*Proof:* Let’s assume that $X \sim f_X(x)$ and $Y \sim f_Y(y)$, then

$$\text{E}(g(X)h(Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f_{X,Y}(x,y) \, dx \, dy$$

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

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

$$= \text{E}(g(X)) \text{E}(h(Y))$$

- $X$ and $Y$ are said to be *uncorrelated* if $\text{Cov}(X, Y) = 0$, or equivalently $\text{E}(XY) = \text{E}(X) \text{E}(Y)$
From our independence result, if \( X \) and \( Y \) are independent then they are uncorrelated.

To show this, set \( g(X) = (X - E(X)) \) and \( h(Y) = (Y - E(Y)) \), then

\[
\text{Cov}(X, Y) = \text{E}[(X - E(X))(Y - E(Y))] = E(X - E(X))E(Y - E(Y)) = 0
\]

However, if \( X \) and \( Y \) are uncorrelated they are not necessarily independent.

Example: Let \( X, Y \in \{-2, -1, 1, 2\} \) such that

\[
p_{X,Y}(1, 1) = \frac{2}{5}, \quad p_{X,Y}(-1, -1) = \frac{2}{5}
\]

\[
p_{X,Y}(-2, 2) = \frac{1}{10}, \quad p_{X,Y}(2, -2) = \frac{1}{10},
\]

\[
p_{X,Y}(x, y) = 0, \text{ otherwise}
\]

Are \( X \) and \( Y \) independent? Are they uncorrelated?

Solution:

Clearly \( X \) and \( Y \) are not independent, since if you know the outcome of one, you completely know the outcome of the other. Let's check their covariance.

\[
\text{E}(X) = \frac{2}{5} - \frac{2}{5} - \frac{2}{10} + \frac{2}{10} = 0, \text{ also}
\]

\[
\text{E}(Y) = 0, \text{ and}
\]

\[
\text{E}(XY) = \frac{2}{5} + \frac{2}{5} - \frac{4}{10} - \frac{4}{10} = 0
\]

Thus, \( \text{Cov}(X, Y) = 0 \), and \( X \) and \( Y \) are uncorrelated!
The Correlation Coefficient

- The correlation coefficient of $X$ and $Y$ is defined as
  $$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

- Fact: $|\rho_{X,Y}| \leq 1$. To show this consider
  $$\mathbb{E} \left[ \left( \frac{X - \mathbb{E}(X)}{\sigma_X} \pm \frac{Y - \mathbb{E}(Y)}{\sigma_Y} \right)^2 \right] \geq 0$$
  $$\frac{\mathbb{E}[(X - \mathbb{E}(X))^2]}{\sigma_X^2} + \frac{\mathbb{E}[(Y - \mathbb{E}(Y))^2]}{\sigma_Y^2} \pm 2 \frac{\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]}{\sigma_X \sigma_Y} \geq 0$$
  $$1 + 1 \pm 2\rho_{X,Y} \geq 0 \implies -2 \leq 2\rho_{X,Y} \leq 2 \implies |\rho_{X,Y}| \leq 1$$

- From the proof, $\rho_{X,Y} = \pm 1$ iff $\frac{X - \mathbb{E}(X)}{\sigma_X} = \pm \frac{Y - \mathbb{E}(Y)}{\sigma_Y}$ (equality with probability 1), i.e., iff $X - \mathbb{E}(X)$ is a linear function of $Y - \mathbb{E}(Y)$

- In general $\rho_{X,Y}$ is a measure of how closely $(X - \mathbb{E}(X))$ can be approximated or estimated by a linear function of $(Y - \mathbb{E}(Y))$

Application: Linear MSE Estimation

- Consider the following signal processing problem:
  $$X \quad \text{Noisy Channel} \quad Y \quad \text{Estimator} \quad \hat{X}$$
  $$aX + b$$

- Here $X$ is a signal (music, speech, image) and $Y$ is a noisy observation of $X$ (output of a noisy communication channel or a noisy circuit). Assume we know the means, variances and covariance of $X$ and $Y$

- Observing $Y$, we wish to find a linear estimate of $X$ of the form $\hat{X} = aY + b$, which minimizes the mean square error
  $$\text{MSE} = \mathbb{E} \left[ (X - \hat{X})^2 \right]$$

- We denote the best such estimate as the minimum mean square estimate (MMSE)
• The MMSE linear estimate of $X$ given $Y$ is given by

$$
\hat{X} = \frac{\text{Cov}(X, Y)}{\sigma_Y^2}(Y - E(Y)) + E(X)
$$

$$
= \rho_{X,Y} \sigma_X \left( \frac{Y - E(Y)}{\sigma_Y} \right) + E(X)
$$

and its MSE is given by

$$
\text{MSE} = \sigma_X^2 - \frac{\text{Cov}^2(X, Y)}{\sigma_Y^2} = (1 - \rho_{X,Y}^2) \sigma_X^2
$$

• Properties of MMSE linear estimate:

  - $E(\hat{X}) = E(X)$, i.e., estimate is unbiased
  - If $\rho_{X,Y} = 0$, i.e., $X$ and $Y$ are uncorrelated, then $\hat{X} = E(X)$ (ignore the observation $Y$)
  - If $\rho_{X,Y} = \pm 1$, i.e., $X - E(X)$ and $Y - E(Y)$ are linearly dependent, then the linear estimate is perfect

Proof

• We first show that $\min_a E \left[ (X - b)^2 \right] = \text{Var}(X)$ and is achieved for $b = E(X)$, i.e., in the absence of any observations, the mean of $X$ is its minimum MSE estimate, and the minimum MSE is $\text{Var}(X)$

To show this consider

$$
E \left[ (X - c)^2 \right] = E \left[ ((X - E(X)) + (E(X) - b))^2 \right]
$$

$$
= E \left[ (X - E(X))^2 \right] + (E(X) - b)^2 + 2(E(X) - b) E(X - E(X))
$$

$$
= E \left[ (X - E(X))^2 \right] + (E(X) - b)^2
$$

$$
\geq E \left[ (X - E(X))^2 \right],
$$

with equality iff $b = E(X)$

• Now, back to our problem. Suppose $a$ has already been chosen. What should $b$ be to minimize $E \left[ (X - aY - b)^2 \right]$? From the above result, we should choose
\[ b = E(X - aY) = E(X) - aE(Y) \]

So, we want to choose \( a \) to minimize

\[ E \left[ ((X - aY) - E(X - aY))^2 \right], \]

which is the same as

\[ E \left[ ((X - E(X)) - a(Y - E(Y)))^2 \right] = \sigma_X^2 + a^2 \sigma_Y^2 - 2a \text{Cov}(X, Y) \]

This is a quadratic function of \( a \). It is minimized when its derivative equals 0, which gives

\[ a = \frac{\text{Cov}(X, Y)}{\sigma_Y^2} = \frac{\rho_{X,Y} \sigma_X \sigma_Y}{\sigma_Y^2} = \frac{\rho_{X,Y} \sigma_X}{\sigma_Y} \]

The mean square error is given by

\[ \sigma_X^2 + a^2 \sigma_Y^2 - 2a \text{Cov}(X, Y) = \sigma_X^2 + \frac{\rho_{X,Y}^2 \sigma_X^2 \sigma_Y^2}{\sigma_Y^2} - 2 \frac{\rho_{X,Y} \sigma_X}{\sigma_Y} \times \rho_{X,Y} \sigma_X \sigma_Y \]

\[ = (1 - \rho_{X,Y}^2) \sigma_X^2 \]

---

### Mean and Variance of Sum of RVs

- Let \( X_1, X_2, \ldots, X_n \) be r.v.s, then by linearity of expectation, the expected value of their sum \( Y \) is

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

Example: **Mean of Binomial r.v.** One way to define a binomial r.v. is as follows:

Flip a coin with bias \( p \) independently \( n \) times and define the Bernoulli r.v.

\( X_i = 1 \) if the \( i \)th flip is a head and \( 0 \) if it is a tail. Let \( Y = \sum_{i=1}^{n} X_i \). Then \( Y \) is a binomial r.v.

Thus

\[ E(Y) = \sum_{i=1}^{n} E(X_i) = np \]

Note that we do not need independence for this result to hold, i.e., the result holds even if the coin flips are not independent (\( Y \) is not binomial in this case, but the expectation doesn’t change)
• Let’s compute the variance of $Y = \sum_{i=1}^{n} X_i$

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

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

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

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

\[
= \sum_{i=1}^{n} \sum_{j=1}^{n} E[(X_i - E(X_i))(X_j - E(X_j))]
\]

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

• If the r.v.s are independent, then $\text{Cov}(X_i, X_j) = 0$ for all $i \neq j$, and

\[
\text{Var}(Y) = \sum_{i=1}^{n} \text{Var}(X_i)
\]

Note that this result only requires that $\text{Cov}(X_i, X_j) = 0$, for all $i \neq j$, and therefore it requires that the r.v.s be uncorrelated (which is in general weaker than independence)

• Example: Variance of Binomial r.v. Again express $Y = \sum_{i=1}^{n} X_i$, where the $X_i$s are i.i.d. Bern($p$). Since the $X_i$s are independent, $\text{Cov}(X_i, X_j) = 0$, for all $i \neq j$. Thus

\[
\text{Var}(Y) = \sum_{i=1}^{n} \text{Var}(X_i) = n \times p(1 - p)
\]

• Example: Hat problem Suppose $n$ people throw their hats in a box and then each picks one hat at random. Let $N$ be the number of people that get back their own hat. Find $E(N)$ and $\text{Var}(N)$

Solution: Define the r.v. $X_i = 1$ if a person selects her own hat, and $X_i = 0$, otherwise. Thus $N = \sum_{i=1}^{n} X_i$
To find the mean and variance of $N$, we first find the means, variances and covariances of the $X_i$s. Note that $X_i \sim \text{Bern}(1/n)$ and thus $E(X_i) = 1/n$, and $\text{Var}(X_i) = (1/n)(1 - 1/n)$. To find the covariance of $X_i$ and $X_j$, $i \neq j$, note that

$$p_{X_i,X_j}(1,1) = \frac{1}{n(n-1)}$$

Thus,

$$\text{Cov}(X_i, X_j) = E(X_iX_j) - E(X_i)E(X_j) = \frac{1}{n(n-1)} \times 1 - \left(\frac{1}{n}\right)^2 = \frac{1}{n^2(n-1)}$$

The mean and variance of $N$ are given by

$$E(N) = nE(X_1) = 1$$

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

$$= n \times \text{Var}(X_1) + n(n-1)\text{Cov}(X_1, X_2)$$

$$= (1 - 1/n) + n(n-1) \times \frac{1}{n^2(n-1)} = 1$$

**Method of Indicators**

- In the last two examples we used the *method of indicators* to simplify the computation of expectation.
- In general the *indicator* of an event $A \subset \Omega$ is a r.v. defined as

$$I_A(\omega) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

Thus

$$E(I_A) = 1 \times P(A) + 0 \times P(A^c) = P(A)$$

The method of indicators involves expressing a given r.v. $Y$ as a sum of indicators in order to simplify the computation of its expectation (this is precisely what we did in the last two examples).

Example: *Spaghetti.* Consider a ball of $n$ spaghetti strands. You randomly pick two strand ends and join them. The process is continued until there are no ends left. Let $X$ be the number of spaghetti loops formed. What is $E(X)$?
Conditional Expectation

- **Conditioning on an event:** Let $X \sim p_X(x)$ be a r.v. and $A$ be a nonzero probability event. We can define the conditional pmf of $X$ given $X \in A$ as

$$p_{X|A}(x) = \frac{P\{X = x, X \in A\}}{P\{X \in A\}} = \begin{cases} \frac{p_X(x)}{P\{X \in A\}}, & \text{if } x \in A \\ 0, & \text{otherwise} \end{cases}$$

Note that $p_{X|A}(x)$ is a pmf on $X$

- Similarly for $X \sim f_X(x)$,

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{P\{X \in A\}}, & \text{if } x \in A \\ 0, & \text{otherwise} \end{cases}$$

is a pdf on $X$

- Example: Let $X \sim \text{Exp}(\lambda)$ and $A = \{X > a\}$, for some constant $a > 0$. Find the conditional pdf of $X$ given $A$

- We define the **conditional expectation** of $g(X)$ given $X \in A$ as

$$E(g(X)|A) = \int_{-\infty}^{\infty} g(x)f_{X|A}(x) \, dx$$

- Example: Find $E(X|A)$ and $E(X^2|A)$ for the previous example.

- **Total expectation:** Let $X \sim f_X(x)$ and $A_1, A_2, \ldots, A_n \subset (-\infty, \infty)$ be disjoint nonzero probability events with $P\{X \in \bigcup_{i=1}^{n} A_i\} = \sum_{i=1}^{n} P\{X \in A_i\} = 1$, then

$$E(g(X)) = \sum_{i=1}^{n} P\{X \in A_i\} E(g(X)|A_i)$$

This is called the **total expectation theorem** and is useful in computing expectation by **divide-and-conquer**

Proof: First note that by the law of total probability

$$f_X(x) = \sum_{i=1}^{n} P\{X \in A_i\} f_{X|A_i}(x)$$
Therefore

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

\[ = \int_{-\infty}^{\infty} g(x) \sum_{i=1}^{n} P\{X \in A_i\} f_{X|A_i}(x) \, dx \]

\[ = \sum_{i=1}^{n} P\{X \in A_i\} \int_{-\infty}^{\infty} g(x) f_{X|A_i}(x) \, dx = \sum_{i=1}^{n} P\{X \in A_i\} E(g(X)|A_i) \]

- **Example: mean and variance of Piecewise uniform pdf** Let \( X \) be a continuous r.v. with the piecewise uniform pdf

\[
 f_X(x) = \begin{cases} 
 1/3, & \text{if } 0 \leq x \leq 1 \\
 2/3, & \text{if } 1 < x \leq 2 \\
 0, & \text{otherwise} 
\end{cases}
\]

Find the mean and variance of \( X \)

**Solution:** Define the events \( A_1 = \{X \in [0, 1]\} \) and \( A_2 = \{X \in (1, 2]\} \)

Then, \( A_1, A_2 \) are disjoint and the sum of their probabilities is 1. The mean of \( X \) can be expressed as

\[
 E(X) = \sum_{i=1}^{2} P\{X \in A_i\} E(X|A_i) = \frac{1}{3} \times \frac{1}{2} + \frac{2}{3} \times \frac{3}{2} = \frac{7}{6}
\]

Also

\[
 E(X^2) = \sum_{i=1}^{2} P\{X \in A_i\} E(X^2|A_i) = \frac{1}{3} \times \frac{1}{3} + \frac{2}{3} \times \frac{7}{3} = \frac{15}{9}
\]

Thus

\[
 \text{Var}(X) = E(X^2) - (E(X))^2 = \frac{11}{36}
\]

- **Mean and variance of a mixed r.v.** As we discussed, there are r.v.s that are neither discrete nor continuous. How do we define their expectation?
Answer: We can express a mixed r.v. $X$ as a mixture of a discrete r.v. $Y$ and a continuous r.v. $Z$ as follows.

Assume that the cdf of $X$ is discontinuous over the set $\mathcal{Y}$ and that $\sum_{y \in \mathcal{Y}} P\{X = y\} = p$. Define the discrete r.v. $Y \in \mathcal{Y}$ to have the pmf

$$p_Y(y) = \frac{1}{p} P\{X = y\}, \ y \in \mathcal{Y}$$

Define the continuous r.v. $Z$ such that

$$f_Z(z) = \begin{cases} \frac{1}{1-p} \frac{dF_X}{dx} & z \notin \mathcal{Y} \\ f_Z(z^-) & z \in \mathcal{Y} \end{cases}$$

Now, we can express $X$ as

$$X = \begin{cases} Y & \text{with probability } p \\ Z & \text{with probability } 1 - p \end{cases}$$

To find $E(X)$, we use the law of total expectation

$$E(X) = p E(X|X \in \mathcal{Y}) + (1 - p) E(X|X \notin \mathcal{Y}) = p E(Y) + (1 - p) E(Z)$$

Now both $E(Y)$ and $E(Z)$ can be computed in the usual way.

---

**Conditioning on a RV**

- Let $(X, Y) \sim f_{X,Y}(x, y)$. If $f_Y(y) \neq 0$, the *conditional pdf* of $X$ given $Y = y$ is given by

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

- We know that $f_{X|Y}(x|y)$ is a pdf for $X$ (function of $y$), so we can define the expectation of any function $g(X, Y)$ w.r.t. $f_{X|Y}(x|y)$ as

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

- If $g(X, Y) = X$, then, the conditional expectation of $X$ given $Y = y$ is

$$E(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) \, dx$$
Example: Let
\[ f_{X,Y}(x, y) = \begin{cases} 2 & \text{for } x, y \geq 0, x + y \leq 1 \\ 0 & \text{otherwise} \end{cases} \]

Find \( E(X|Y = y) \) and \( E(XY|Y = y) \)

Solution: We already know that
\[ f_{X|Y}(x|y) = \begin{cases} \frac{1}{1-y} & \text{for } x, y \geq 0, x + y \leq 1, y < 1 \\ 0 & \text{otherwise} \end{cases} , \]

thus
\[ E(X|Y = y) = \int_0^{1-y} \frac{1}{1-y} x \, dx = \frac{1-y}{2} \text{ for } 0 \leq y < 1 \]

Now to find \( E(XY|Y = y) \), note that
\[ E(XY|Y = y) = y E(X|Y = y) = \frac{y(1-y)}{2} \text{ for } 0 \leq y < 1 \]

Conditional Expectation as a RV

- We define the \textit{conditional expectation} of \( g(X, Y) \) given \( Y \) as the random variable \( E(g(X, Y)|Y) \), which is a function of the random variable \( Y \)

- So, \( E(X|Y) \) is the conditional expectation of \( X \) given \( Y \), a r.v. that is a function of \( Y \)

- Example: This is a continuation of the previous example. Find the pdf of \( E(X|Y) \)

Solution: The conditional expectation of \( X \) given \( Y \) is the r.v.
\[ E(X|Y) = \frac{1-Y}{2} \triangleq Z \]

The pdf of \( Z \) is given by
\[ f_Z(z) = 8z \text{ for } 0 < z \leq \frac{1}{2} \]
Now let’s find the expected value of the r.v. \( Z \)

\[
E(Z) = \int_{0}^{1/2} 8z^2 \, dz = \frac{1}{3} = E(X)
\]

**Iterated Expectation**

- In general we can find \( E(g(X, Y)) \) using *iterated expectation* as

\[
E(g(X, Y)) = E_Y \left[ E_X(g(X, Y) | Y) \right],
\]

where \( E_X \) means expectation w.r.t. \( f_{X|Y}(x|y) \) and \( E_Y \) means expectation w.r.t. \( f_Y(y) \). To show this consider

\[
E_Y \left[ E_X(g(X, Y) | Y) \right] = \int_{-\infty}^{\infty} E_X(g(X, Y) | Y = y) f_Y(y) \, dy
\]

\[
= \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x|y) \, dx \right) f_Y(y) \, dy
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) \, dx \, dy
\]

\[
= E(g(X, Y))
\]

- This result can be very useful in computing expectation
Example: A coin has random bias \( P \) with
\[
f_P(p) = 2(1 - p)
\]
for \( 0 \leq p \leq 1 \). The coin is flipped \( n \) times independently. Let \( N \) be the number of heads, find \( E(N) \)
Solution: Of course we can first find the pmf of \( N \), then find its expectation. Using iterated expectation we can find \( N \) faster. Consider
\[
E(N) = E_P[E_N(N|P)] = E_P(nP) = n \int_0^1 2(1 - p)p \, dp = \frac{n}{3}
\]

Example: Let \( E(X|Y) = Y^2 \), and \( Y \sim U[0, 1] \), find \( E(X) \)
Solution: Here we cannot first find the pdf of \( X \), since we do not know \( f_{X|Y}(x|y) \), but using iterated expectation we can easily find
\[
E(X) = E_Y[E_X(X|Y)] = \int_0^1 y^2 \, dy = \frac{1}{3}
\]

### Conditional Variance

Let \( X \) and \( Y \) be two r.v.s. We define the conditional variance of \( X \) given \( Y = y \) as
\[
\text{Var}(X|Y = y) = E[(X - E(X|Y = y))^2|Y = y] = E(X^2|Y = y) - [E(X|Y = y)]^2
\]

The r.v. \( \text{Var}(X|Y) \) is simply a function of \( Y \) that takes on the values \( \text{Var}(X|Y = y) \). Its expected value is
\[
E_Y[\text{Var}(X|Y)] = E_Y[E(X^2|Y) - (E(X|Y))^2] = E(X^2) - E[(E(X|Y))^2]
\]
Since \( E(X|Y) \) is a r.v., it has a variance
\[
\text{Var}(E(X|Y)) = E_Y[\text{Var}(E(X|Y))] = E[(E(X|Y) - E[E(X|Y)])^2] = E[(E(X|Y))^2] - (E(X))^2
\]

Law of Conditional Variances: We can show that
\[
\text{Var}(X) = E(\text{Var}(X|Y)) + \text{Var}(E(X|Y))
\]
Proof: Simply add above expressions for \( E(\text{Var}(X|Y)) \) and \( \text{Var}(E(X|Y)) \):
\[
E(\text{Var}(X|Y)) + \text{Var}(E(X|Y)) = E(X^2) - (E(X))^2 = \text{Var}(X)
\]
Application: Nonlinear MSE Estimation

- Consider the estimation setup with signal $X$, observation $Y$
- Assume we know the pdf of $X$ and the conditional pdf of the channel $f_{Y|X}(y|x)$ for all $(x, y)$
- We wish to find the best nonlinear estimate $\hat{X} = g(Y)$ of $X$ that minimizes the mean square error

$$\text{MSE} = \mathbb{E}[(X - \hat{X})^2] = \mathbb{E}[(X - g(Y))^2]$$

- The $\hat{X}$ that achieves the minimum MSE is called the minimum MSE estimate (MMSE) of $X$ (given $Y$)

**MMSE Estimate**

- The MMSE estimate of $X$ given the observation $Y$ and complete knowledge of the joint pdf $f_{X,Y}(x,y)$ is

$$\hat{X} = \mathbb{E}(X|Y)$$

and its MSE (i.e., minimum MSE) is

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

- Proof:
  - Recall that $\min_a \mathbb{E}[(X - a)^2] = \text{Var}(X)$ and is achieved for $a = \mathbb{E}(X)$, i.e., in the absence of any observations, the mean of $X$ is its minimum MSE estimate, and the minimum MSE is $\text{Var}(X)$
  - We now use this result to show that $\mathbb{E}(X|Y)$ is the MMSE estimate of $X$ given $Y$. First we use iterated expectation to write
\[ E[(X - g(Y))^2] = E_Y[E_X[(X - g(Y))^2|Y]] \]

From the above result we know that for each \( Y = y \),
\( E_X[(X - g(y))^2|Y = y] \) is minimum for \( g(y) = E(X|Y = y) \). Thus the
MSE is minimized for \( g(Y) = E(X|Y) \).
Thus \( E(X|Y) \) minimizes the MSE conditioned on every \( Y = y \) and not just
its average over \( Y \)!

- **Properties of the minimum MSE estimator:**
  - Since \( E(\hat{X}) = E[E(X|Y)] = E(X) \), the best MSE estimate is called *unbiased*.
  - If \( X \) and \( Y \) are independent, then the best MSE estimate is \( E(X) \).
  - The conditional expectation of the estimation error \( E((X - \hat{X})|Y = y) = 0 \)
    for all \( y \), i.e., the error is unbiased for every \( Y = y \).
  - From the law of conditional variance
    \[ \text{Var}(X) = \text{Var}(\hat{X}) + E(\text{Var}(X|Y)) \],
    i.e., the sum of the variance of the estimate and the minimum MSE is equal
to the variance of the signal.
Example

- Again let

\[ f(x, y) = \begin{cases} 
2 & \text{for } x, y \geq 0, \ x + y \leq 1 \\
0 & \text{otherwise}
\end{cases} \]

Find the MMSE estimate of \( X \) given \( Y \) and its MSE

Solution: We already found the MMSE estimate to be

\[ E(X|Y) = \frac{1 - Y}{2} \text{ for } 0 \leq Y \leq 1 \]

We know that

\[ f_{X|Y}(x|y) = \begin{cases} 
1/(1 - y) & \text{for } x, y \geq 0, \ x + y \leq 1, \ y < 1 \\
0 & \text{otherwise}
\end{cases} \]

Thus for \( Y = y \), the minimum MSE is given by

\[ \text{Var}(X|Y = y) = \frac{(1 - y)^2}{12} \text{ for } 0 \leq y < 1 \]

Thus the minimum MSE is \( E_Y(\text{Var}(X|Y)) = 1/24 \), compared to \( \text{Var}(X) = 1/18 \). By the law of conditional variance, the difference is the variance of the estimate \( \text{Var}(E(X|Y)) = 1/18 - 1/24 = 1/72 \)
Sum of Random Number of Independent RVs

• Let $N$ be a r.v. taking positive integer values and $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.s independent of $N$

• Define the sum

$$Y = \sum_{i=1}^{N} X_i$$

• We are given $E(N)$, $\text{Var}(N)$, the mean of $X_i$, $E(X)$, and its variance $\text{Var}(X)$ and wish to find the mean and variance of $Y$

• Using iterated expectation, the mean is:

$$E(Y) = E_N \left[ E \left( \sum_{i=1}^{N} X_i | N \right) \right]$$

$$= E_N \left( \sum_{i=1}^{N} E(X_i | N) \right) \text{ by linearity of expectation}$$

$$= E_N \left( \sum_{i=1}^{N} E(X_i) \right) \quad X_i \text{ and } N \text{ are independent}$$

$$= E_N [N E(X)] = E(N) E(X)$$

Using the law of conditional variance, the variance is:

$$\text{Var}(Y) = E \left[ \text{Var}(Y | N) + \text{Var}(E(Y | N)) \right]$$

$$= E \left[ N \text{Var}(X) \right] + \text{Var}(N E(X))$$

$$= \text{Var}(X) E(N) + (E(X))^2 \text{Var}(N)$$

• Example: You visit bookstores looking for a copy of the *Great Expectations*.
Each bookstore carries the book with probability $p$, independent of all other bookstores. You keep visiting bookstores until you find the book. In each bookstore visited, you spend a random amount of time, exponentially distributed with parameter $\lambda$. Assuming that you will keep visiting bookstores until you buy the book and that the time spent in each is independent of everything else, find the mean and variance of the total time spent in bookstores.

Solution: The total number of bookstores visited is a r.v. $N \sim \text{Geom}(p)$. Let $X_i$ be the amount of time spent in each bookstore. Thus, $X_1, X_2, \ldots$ are i.i.d. Exp($\lambda$) r.v.s. Now let $Y$ be the total amount of time spent looking for the books. Then

$$Y = \sum_{i=1}^{N} X_i$$

The mean and variance of $Y$, thus, are

$$E(Y) = E(N) E(X) = \frac{1}{p\lambda}$$

$$\text{Var}(Y) = \text{Var}(X) E(N) + (E(X))^2 \text{Var}(N)$$

$$= \frac{1}{p\lambda^2} + \frac{1}{\lambda^2} \times \frac{1-p}{p^2} = \frac{1}{p^2\lambda^2}$$
Lecture Notes 5
Transforms

- Definition
- Examples
- Moment Generating Properties
- Inversion
- Transforms of Sums of Independent RVs


Definition

- The transform associated with the probability distribution of a random variable $X$ is defined as

$$M_X(s) = E(e^{sX})$$

$$= \begin{cases} 
\sum_x e^{sx}p_X(x), & \text{for } X \sim p_X(x) \\
\int_{-\infty}^{\infty} e^{sx}f_X(x) \, dx, & \text{for } X \sim f_X(x),
\end{cases}$$

where $s$ is a scalar parameter

- Closely related to Fourier, Laplace and $z$ transforms (depends on allowed values of $s$)

- When we use $s$, the transform is called the moment generating function. When we plug $s = i\omega$, where $i = \sqrt{-1}$ and $\omega$ is a real-valued parameter, the transform is called the characteristic function

- Important fact: There is a one-to-one correspondence between the moment generating function for a r.v. and its probability distribution (pmf or pdf)
Moment Generating Function: Examples

- \( X \sim \text{Bern}(p) \): \( p_X(1) = p, \ p_X(0) = 1 - p \)

\[
M_X(s) = E(e^{sX}) = \sum_{k=0}^{1} e^{sk} p_X(k) = (1 - p) + pe^s
\]

- \( X \sim \text{Geom}(p) \): \( p_X(k) = p(1-p)^{k-1}, \ k = 1, 2, \ldots \)

\[
M_X(s) = E(e^{sX}) = \sum_{k=1}^{\infty} p_X(k)e^{sk} = \sum_{k=1}^{\infty} p(1-p)^{k-1}e^{sk}
\]

\[
= pe^s \sum_{k=0}^{\infty} ((1-p)e^s)^k = \frac{pe^s}{1 - (1-p)e^s},
\]

provided \((1-p)e^s < 1\), which implies that \( s < \ln(1/(1-p)) \)

Note: In general, \( M_X(s) \) is defined only over values of \( s \) where it is finite

- \( X \sim \text{Poisson}(\lambda) \): \( p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \ k \geq 0 \)

\[
M_X(s) = E(e^{sX}) = \sum_{k=0}^{\infty} e^{sk} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{a^k}{k!} \text{ where } a = \lambda e^s
\]

\[
= e^{-\lambda} \times e^a = e^{\lambda(e^s-1)}
\]

- \( X \sim \text{Exp}(\lambda) \): \( f_X(x) = \lambda e^{-\lambda x}, \ x \geq 0 \)

\[
M_X(s) = E(e^{sX}) = \int_{0}^{\infty} \lambda e^{-\lambda x} e^{sx} \, dx = \frac{\lambda}{\lambda - s} \text{ for } s < \lambda
\]
\[ X \sim N(\mu, \sigma^2): \quad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

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

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

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

\[ = e^{s\mu + \frac{s^2 \sigma^2}{2}} \]

In particular, the transform associated with \( X \sim N(0, 1) \) is

\[ M_X(s) = e^{s^2/2} \]

---

**Inversion of Transforms**

- Key fact: The transform uniquely determines the pmf or pdf
- Generally, inversion can be hard, best treated in transform classes
- For this course, inversion is accomplished by by tricks, by inspection, or by looking it up
Why are Transforms Important?

- Alternative description of a probability law (yawn)
- Just like in basic systems and signals, transforming can often greatly simplify computations (which is why Fourier analysis, Laplace transforms, and z transforms are so important)
  - Used to simplify computation of moments ($E(X^k)$)
  - Used to simplify computation of pdf for sum of independent random variables (convolution becomes multiplication)
- Used to prove certain properties of r.v.s
- Used to prove the Central Limit Theorem

Moment Generating Properties

- Let $X \sim p_X(x)$. First differentiate $M_X(s)$ with respect to $s$:

$$M_X^{(1)}(s) \triangleq \frac{d}{ds}M_X(s)$$

$$= \frac{d}{ds} \sum_x p_X(x)e^{sx}$$

$$= \sum_x xp_X(x)e^{sx}$$

Now, evaluating the derivative at $s = 0$, we find that

$$M_X^{(1)}(0) = \frac{d}{ds}M_X(s) \bigg|_{s=0} = E(X)$$

Thus the mean of a random variable can be found by differentiating the moment generating function and setting the argument to 0 as

$$E(X) = M_X^{(1)}(0)$$

Same result holds for continuous r.v.s
• Repeated differentiation:

\[ M^{(k)}(0) \triangleq \frac{d^k}{ds^k}M_X(s) \Big|_{s=0} = \int_{-\infty}^{\infty} x^k f_X(x) \, dx = E(X^k) \]

• Examples:
  ◦ \( X \sim \text{Bern}(p) \) has transform \( M_X(s) = (1 - p) + pe^s \). Hence
    \[
    E(X) = M_X^{(1)}(0) = p, \quad \text{and} \quad E(X^2) = M_X^{(2)}(0) = p
    \]
    In fact, for any \( k \geq 1 \), \( E(X^k) = M_X^{(k)}(0) = p \)
  ◦ \( X \sim \text{Exp}(\lambda) \) has transform \( M_X(s) = \lambda/\lambda - s \)
    \[
    E(X) = M_X^{(1)}(0) \quad = \frac{\lambda}{(\lambda - s)^2} \bigg|_{s=0} = \frac{1}{\lambda}
    \]
    \[
    E(X^2) = M_X^{(2)}(0) \quad = \frac{2 \lambda}{(\lambda - s)^3} \bigg|_{s=0} = \frac{2}{\lambda^2}
    \]
  ◦ \( X \sim \mathcal{N}(\mu, \sigma^2) \) has transform \( M_X(s) = e^{s\mu + s^2\sigma^2/2} \)
    \[
    E(X) = M_X^{(1)}(0) \quad = (\mu + s\sigma^2)e^{s\mu + s^2\sigma^2/2} \bigg|_{s=0} = \mu
    \]
    \[
    E(X^2) = M_X^{(2)}(0) \quad = \left( (\mu + s\sigma^2)^2 e^{s\mu + s^2\sigma^2/2} + \sigma^2 e^{s\mu + s^2\sigma^2/2} \right) \bigg|_{s=0}
    \]
    \[
    = \sigma^2 + \mu^2
    \]
    In general, can show that for \( X \sim \mathcal{N}(0, \sigma^2) \):
    \[
    E\left( X^k \right) = M_X^{(k)}(0) = \begin{cases} 
    1 \times 3 \times \ldots \times (k-1) \sigma^k, & \text{if } k \text{ is even} \\
    0, & \text{otherwise}
    \end{cases}
    \]
## Moment Generating Function for Famous Random Variables

<table>
<thead>
<tr>
<th>Random Variable</th>
<th>PMF/PDF</th>
<th>$M(s)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Bern}(p)$</td>
<td>$p_X(1) = p$, $p_X(0) = 1 - p$</td>
<td>$pe^s + (1 - p)$</td>
</tr>
<tr>
<td>$\text{Geom}(p)$</td>
<td>$p_X(k) = p(1 - p)^{k-1}$, $k = 1, 2, \ldots$</td>
<td>$\frac{pe^s}{1 - (1-p)e^s}$</td>
</tr>
<tr>
<td>$\text{B}(n, p)$</td>
<td>$p_X(k) = \binom{n}{k}p^k(1-p)^{n-k}$, $k = 0, 1, \ldots, n$</td>
<td>$(pe^s + (1 - p))^n$</td>
</tr>
<tr>
<td>$\text{Poisson}(\lambda)$</td>
<td>$p_X(k) = \frac{\lambda^k}{k!}e^{-\lambda}$, $k = 0, 1, \ldots$</td>
<td>$e^{\lambda(e^s-1)}$</td>
</tr>
<tr>
<td>$\text{Exp}(\lambda)$</td>
<td>$f_X(x) = \lambda e^{-\lambda x}$, $x \geq 0$</td>
<td>$\frac{\lambda}{\lambda - s}$</td>
</tr>
<tr>
<td>$\mathcal{N}(\mu, \sigma^2)$</td>
<td>$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$</td>
<td>$e^{\mu s + \frac{s^2}{2}}$</td>
</tr>
</tbody>
</table>

### Sums of Independent Random Variables

- Suppose that $X$ and $Y$ are independent r.v.s

Recall that for any functions $g$ and $h$,

$$E(g(X)h(Y)) = E(g(X)) E(h(Y))$$

In particular, suppose that $W = X + Y$ and we wish to find $M_W(s)$. Then

$$M_W(s) = E(e^{sW})$$

$$= E(e^{s(X+Y)})$$

$$= E(e^{sX}e^{sY})$$

$$= E(e^{sX}) E(e^{sY}) = M_X(s)M_Y(s)$$

I.e., adding independent r.v.s produces a new r.v. whose transform is the product of the original transforms

Recall that the pdf of $W$, $f_W(w) = f_X(w) * f_Y(w)$. So, this result corresponds to the well known fact that convolutions transform into products
• This result can be extended to more than 2 r.v.s:

Let \( X_1, X_2, \ldots, X_n \) be mutually independent random variables with moment generating functions \( M_{X_1}, M_{X_2}, \ldots, M_{X_n} \), respectively, then the moment generating function of their sum \( Y = \sum_{i=1}^{n} X_i \) is

\[
M_Y(s) = \prod_{i=1}^{n} M_{X_i}(s)
\]

If the \( X_i \)s are independent and identically distributed with a common moment generating function \( M_X(s) \), then

\[
M_Y(s) = (M_X(s))^n
\]

• So to find the pdf of a sum of independent r.v.s you can either convolve their pdfs or multiply their moment generating functions and then invert the result.

EE 178: Transforms

**Examples**

• Moment generating function for a Binomial r.v.: Let \( Y \sim B(n, p) \). We wish to find its moment generating function \( M_Y(s) \)

Recall that we can express \( Y \) as a sum of \( n \) i.i.d. Bern\((p)\) r.v.s, \( X_1, X_2, \ldots, X_n \). Since the moment generating function of a Bern\((p)\) r.v. is

\[
M_X(s) = \sum_{k=0}^{1} e^{sk} p_X(k) = (1 - p) + pe^s,
\]

the moment generating function of the binomial r.v. \( Y = \sum_{i=1}^{n} X_i \), is

\[
M_Y(s) = ((1 - p) + pe^s)^n
\]

To verify, let’s find the inverse. By the Newton expansion formula

\[
((1 - p) + pe^s)^n = \sum_{k=0}^{n} \left( \binom{n}{k} p^k (1 - p)^{n-k} \right) e^{sk}
\]

\[
= \sum_{k=0}^{n} p_Y(k) e^{sk}
\]
Suppose we have two independent Gaussian r.v.s, $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and we wish to find the pdf of their sum $W = X + Y$

We first find the transform of their sum

$$M_W(s) = M_X(s)M_Y(s)$$

$$= e^{s\mu_X + s^2\sigma_X^2/2}s\mu_Y + s^2\sigma_Y^2/2$$

$$= e^{s(\mu_X + \mu_Y) + s^2(\sigma_X^2 + \sigma_Y^2)/2}$$

We now invert the transform to find the pdf. This is very easy since the form of the transform is that of a Gaussian pdf with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$, i.e., $W \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

So, if we add two (or more) independent Gaussian r.v.s, we obtain a Gaussian r.v.. This shows that Gaussian distribution is infinitely divisible (see Lecture Notes 3)

---

**Distribution of a Random Sum of RVs**

- Let $N$ be a r.v. taking positive integer values and $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.s independent of $N$

- Define the sum

$$Y = \sum_{i=1}^{N} X_i$$

- Using conditional expectation, we already found the mean and variance of $Y$ in terms of the mean and variance of $X$ and $N$

- Now suppose we are given the transforms for $X$, $M_X(s)$, and $N$, $M_N(s)$, and we wish to find the transform of $Y$. We again use conditional expectation

$$E(e^{sY} | N = n) = E(e^{sX_1}e^{sX_2}\ldots e^{sX_N} | N = n)$$

$$= E(e^{sX_1}e^{sX_2}\ldots e^{sX_n})$$

$$= E(e^{sX_1})E(e^{sX_2})\ldots E(e^{sX_n})$$

$$= (M_X(s))^n$$
Using iterated expectation

\[ M_Y(s) = E \left[ E (e^{sY} \mid N) \right] \]
\[ = E \left[ (M_X(s))^N \right] \]

Note the similarity between \( M_Y(s) \) above and \( M_N(s) = E (e^{sN}) \)

They are the same except that \( M_Y \) replaces \( e^s \) in \( M_N \) by \( M_X(s) \). In other words, if we know the functional form of \( M_N(s) \) and it depends on \( s \) only through \( e^s \), then we can find \( M_Y \) by simply replacing \( e^s \) by \( M_X(s) \).

- Example: Recall the bookstore problem. \( N \sim \text{Geom}(p) \) is the number of stores visited and \( X_1, X_2, \ldots \) are i.i.d. \( \text{Exp}(\lambda) \) r.v.s and represent the times spent in each store. The total amount of time spent looking for the book until it’s found is

\[ Y = \sum_{i=1}^{N} X_i \]

Find the pdf of \( Y \)

Solution: We find the transform associated with \( Y \) and then invert it to find the pdf. First note that the transform for \( N \) is

\[ M_N(s) = \frac{pe^s}{1 - (1 - p)e^s} \]

Thus, the transform for \( Y \) is

\[ M_Y(s) = E \left[ (M_X(s))^N \right] \]
\[ = \frac{pM_X(s)}{1 - (1 - p)M_X(s)} \]

But, since \( X \sim \text{Exp}(\lambda) \), its transform is

\[ M_X(s) = \frac{\lambda}{\lambda - s} \]

Substituting, we obtain

\[ M_Y(s) = \frac{p\left(\frac{\lambda}{\lambda - s}\right)}{1 - (1 - p)\left(\frac{\lambda}{\lambda - s}\right)} = \frac{p\lambda}{p\lambda - s} \]

Thus, \( Y \sim \text{Exp}(p\lambda) \)!
Lecture Notes 6
Limit Theorems

- Motivation
- Markov and Chebyshev Inequalities
- Weak Law of Large Numbers
- The Central Limit Theorem
- Confidence Intervals


Motivation

- One of the key questions in statistics is how to estimate the statistics of a r.v., e.g., its mean, variance, distribution, etc.

- To estimate such a statistic, we collect samples and use an estimator in the form of a sample average
  - How good is the estimator? Does it “converge” to the true statistic?
  - How many samples do we need to ensure with some confidence that we are within a certain range of the true value of the statistic?

- These questions are answered using probability theory. The answers are called limit theorems

- Example: Suppose we are given a coin with unknown bias $p$. To estimate the bias we flip the coin $n$ times and compute the relative frequency of occurrence of heads $n_H/n$.
  - Does $n_H/n \to p$?
  - How many coin flips do we need?
The Sample Mean

• We can cast the above example as an example of estimating the mean of a r.v.. Suppose \( X \) is a r.v. with finite but unknown mean \( E(X) \) (e.g., for the coin flipping example \( X \sim \text{Bern}(p) \) and we want to estimate \( E(X) = p \)).

• To estimate the mean, we collect \( X_1, X_2, \ldots, X_n \) i.i.d. samples drawn according to the same distribution as \( X \) and compute the sample mean

\[
S_n = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

○ Does \( S_n \rightarrow E(X) \) as \( n \rightarrow \infty \)?
○ What do we mean by convergence here?
○ How large should \( n \) be?

Mean and variance of \( S_n \): From previous discussions we know that

\[
E(S_n) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = E(X)
\]

\[
\text{Var}(S_n) = \frac{1}{n^2} \text{Var} \left( \sum_{i=1}^{n} X_i \right) = \frac{\sigma_X^2}{n}
\]

• Thus,

\[
\text{Var}(S_n) = E \left[ (S_n - E(S_n))^2 \right] = E \left[ (S_n - E(X))^2 \right] \rightarrow 0, \text{ as } n \rightarrow \infty
\]

• This result is an example of limit theorem. It says that the sample mean converges in mean square to the true mean of the r.v.

• We will prove another limit theorem called the Weak Law of Large Numbers using this result

• We then answer the question of how many samples are needed using the Central Limit Theorem
Using Expectation to Bound Probability

- In many cases we do not know the distribution of a r.v. $X$ but want to find the probability of an event such as \( \{ X > a \} \) or \( \{|X - E(X)| > a\} \)

- The Markov and Chebyshev inequalities can help us obtain bounds on the probabilities of such events in terms of the r.v. mean and variance

- Example: Let $X \geq 0$ represent the age of a person in the Bay Area. Assume that we know the mean $E(X) = 35$ years, what fraction of the population is 70 years old or older?

  Clearly we cannot answer this question knowing only the mean, but we can with certainty say that $P\{X \geq 70\} \leq 0.5$, since otherwise the mean would be larger than 35

- This is an application of the Markov inequality

---

**Markov Inequality**

- For any r.v. $X \geq 0$ with finite mean $E(X)$ and any $a > 1$,

  $$P\{X \geq a E(X)\} \leq \frac{1}{a}$$

- Proof: Define the *indicator function* of the set $A = \{x \geq a E(X)\}$ as

  $$I_A(x) = \begin{cases} 1 & \text{for } x \geq a E(X) \\ 0 & \text{otherwise} \end{cases}$$

  ![Diagram of Markov Inequality](image-url)
From the figure it follows that
\[ I_A(X) \leq \frac{X}{aE(X)} \]

Taking the expectation of both sides, we obtain the Markov inequality
\[ E(I_A(X)) = P\{X \geq aE(X)\} \leq E\left(\frac{X}{aE(X)}\right) = 1/a \]

- The Markov inequality can be very loose. Let \( X \sim \text{Exp}(1) \), then
\[ P\{X \geq 10\} = e^{-10} \approx 4.54 \times 10^{-5} \]

The Markov inequality gives
\[ P\{X \geq 10\} \leq \frac{1}{10}, \]

which is very pessimistic

\[ \text{Chebyshev Inequality} \]

- Let \( X \) be a device parameter in an integrated circuit (IC) with known mean and variance. The IC is out-of-spec if \( X \) is more than, say, \( 3\sigma_X \) away from its mean. We wish to find the fraction of out-of-spec ICs, \( P\{|X - E(X)| \geq 3\sigma_X\} \)
- The Chebyshev inequality gives us an upper bound on this fraction in terms the mean and variance of \( X \)
- Let \( X \) be a r.v. with finite mean \( E(X) \) and variance \( \sigma_X^2 \). The Chebyshev inequality states that for any \( a > 1 \)
\[ P\{|X - E(X)| \geq a\sigma_X\} \leq \frac{1}{a^2} \]

- Proof: We use the Markov inequality. Define the r.v. \( Y = (X - E(X))^2 \geq 0 \). Since \( E(Y) = \sigma_X^2 \), the Markov inequality gives
\[ P\{Y \geq a^2\sigma_X^2\} \leq \frac{1}{a^2} \]
But
\[ \{|X - E(X)| \geq a\sigma_X\} \text{ occurs iff } \{Y \geq a^2\sigma_X^2\} \]

Thus
\[ P\{|X - E(X)| \geq a\sigma_X\} \leq \frac{1}{a^2} \]

- The Chebyshev inequality can be very loose. Let \( X \sim \mathcal{N}(0, 1) \). Using the Chebyshev inequality, we obtain
\[ P\{|X| \geq 3\} \leq \frac{1}{9}, \]
which is very pessimistic compared to the actual value \( \approx 2 \times 10^{-3} \)

---

**The Weak Law of Large Numbers**

- The WLLN states that if \( X_1, X_2, \ldots, X_n \) is a sequence of i.i.d. r.v.s with finite mean \( E(X) \) and variance \( \sigma_X^2 \), then for any \( \epsilon > 0 \),
\[ \lim_{n \to \infty} P\{|S_n - E(X)| > \epsilon\} = 0 \]

- Proof: By the Chebyshev inequality
\[ P\{|S_n - E(S_n)| > \epsilon\} \leq \frac{\text{Var}(S_n)}{\epsilon^2} \]

But, \( E(S_n) = E(X) \) and \( \text{Var}(S_n) = \sigma_X^2/n \), thus
\[ P\{|S_n - E(X)| > \epsilon\} \leq \frac{\sigma_X^2}{n\epsilon^2} \to 0, \text{ as } n \to \infty \]

- Note that the WLLN holds if \( X_1, X_2, \ldots, X_n \) are only pairwise independent, which is weaker than mutual independence.
  In fact, we only need the r.v.s to be uncorrelated, which is in general even weaker than pairwise independence.
Confidence Intervals

• Given \( \epsilon, \delta > 0 \), how large should the number of samples \( n \) be such that

\[
P\{|S_n - E(X)| \leq \epsilon\} \geq 1 - \delta,
\]

i.e., with probability \( \geq (1 - \delta) \), \( S_n \) is within \( \pm \epsilon \) of \( E(X) \)

• Let's use the Chebyshev inequality

\[
P\{|S_n - E(X) - E(S_n)| \leq \epsilon\} = 1 - P\{|S_n - E(S_n)| > \epsilon\} \geq 1 - \frac{\text{Var}(S_n)}{\epsilon^2} = 1 - \frac{\sigma_X^2}{n \epsilon^2}
\]

So if we let \( \sigma_X^2/n \epsilon^2 \leq \delta \), we obtain \( n \geq \sigma_X^2/\delta \epsilon^2 \)

• Example: Let \( \epsilon = 0.1 \sigma_X \) and \( \delta = .001 \), then the number of samples

\[
n \geq \sigma_X^2/10^{-2}\sigma_X^210^{-3} = 10^5,
\]

i.e., \( 10^5 \) samples ensure that \( S_n \) is within \( \pm 0.1 \sigma_X \) of \( E(X) \) with probability (confidence) of 0.999 or better, independent of the distribution of \( X \)

• The Pollster's Problem: Let \( p \) be the fraction of people that intend to vote for Snidely Whiplash against Dudley Doright. A pollester polls \( n \) people.

For the \( i \) th person polled define the r.v.

\[
X_i = \begin{cases} 1 & \text{prefers Whiplash} \\ 0 & \text{otherwise} \end{cases}
\]

Let \( S_n = \frac{1}{n} \sum_{i=1}^{n} X_i \) be the sample mean of the \( X_i \)'s, i.e., the fraction of people polled that prefer Whiplash

The pollster wants to choose \( n \) large enough to ensure that

\[
P\{|S_n - p| \leq .01\} \geq 0.95
\]

• Solution: By the Chebyshev inequality,

\[
P\{|S_n - p| \leq .01\} \geq 1 - \frac{\sigma_X^2}{n(0.01)^2}
\]

But, \( X \sim \text{Bern}(p) \), thus \( \sigma_X^2 = p(1-p) \leq 0.25 \). Thus \( n = 50,000 \) suffices (as we shall see, 50,000 is wildly conservative)
Gaussian Random Variables

- A Gaussian random variable $X \sim N(\mu, \sigma^2)$, has the pdf
  \[
  f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
  \]

- Gaussian has several nice properties:
  - If $Y = aX + b$, then $Y \sim N(a\mu + b, a^2\sigma^2)$
  - The transform closely resembles the form of the pdf: $M_X(s) = e^{s\mu + s^2\sigma^2/2}$
  - If $X_1, X_2, \ldots, X_n$ are independent and $X_i \sim N(\mu_i, \sigma_i^2)$, for $i = 1, 2, \ldots, n$, then $W = \sum_i a_i X_i \sim N\left(\sum_i a_i \mu_i, \sum_i a_i^2 \sigma_i^2\right)$
    In particular, any weighted sum of i.i.d. Gaussian r.v.s is Gaussian

- We shall see that a certain weighted sum of i.i.d. r.v.s is approximately Gaussian even when the r.v.s themselves are not Gaussian — a result known as the Central Limit Theorem (CLT)
  - An important reason for the great popularity of Gaussian models

The Standardized Sum

- Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. r.v.s with finite mean $E(X)$ and variance $\sigma_X^2$.
- We define the “standardized” or “normalized” sum
  \[
  Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left( \frac{X_k - E(X)}{\sigma_X} \right)
  \]
- $Z_n$ is called “standardized” because $E(Z_n) = 0$, and $\sigma_{Z_n}^2 = 1$
- In the special case where $E(X) = 0$ and $\sigma_X^2 = 1$, the standardized sum becomes
  \[
  Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k
  \]
  Note the similarity to the sample average $S_n$ considered for the law of large numbers: $Z_n = \sqrt{n} S_n$
The Central Limit Theorem

- Let $F_{Z_n}(z)$ be the cdf of $Z_n$. The Central Limit Theorem states that

$$\lim_{n \to \infty} F_{Z_n}(z) = 1 - Q(z)$$

- Intuitively, if we sum up a large number of independent random variables and normalize by $n^{-1/2}$, the result is approximately Gaussian. Examples:
  - A voltage meter across a resistor measures the thermal noise voltage, which is the sum of the effects of billions of electrons randomly moving and colliding with each other. Regardless of the probabilistic description of these micro-events, the global noise voltage appears to be Gaussian.
  - Dust particles suspended in water are subjected to the random collisions of millions of molecules. The motion of any individual particle (called “Brownian motion”) in three dimensions appears to be Gaussian.

- Example: Let $X_1, X_2, \ldots$ be i.i.d. $U[-1, 1]$ r.v.s and $Z_n$ be as defined before. The following plots show the pdf of $Z_n$ for $n = 1, 2, 4, 16$. Note how fast the pdf of $Z_n$ becomes close to Gaussian.
Example: Let $X_1, X_2, \ldots$ be i.i.d. $\text{Bern}(1/2)$. Define $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - 0.5}{0.5}$ as before. The following plots show the cdf of $Z_n$ for $n = 10, 20, 160$

Of course $Z_n$ is discrete, thus has no pdf. Its cdf converges to the Gaussian cdf

Proof of The CLT

Let $X_1, X_2, \ldots, X_n$ be i.i.d. with finite mean $E(X)$ and variance $\sigma_X^2$ and define the standardized sum as

$$Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left( \frac{X_k - E(X)}{\sigma_X} \right)$$

Let the transform of the standardized r.v. $Y = (X - E(X))/\sigma_X$ be $M_Y(s)$, then the transform of $Z_n$ can be expressed as

$$M_{Z_n}(s) = \left( M_Y \left( \frac{s}{\sqrt{n}} \right) \right)^n$$

Now, as $n \to \infty$, $s/\sqrt{n} \to 0$. Therefore, provided $M_Y$ is suitably smooth, we can use the Taylor series to obtain the approximation

$$M_Y \left( \frac{s}{\sqrt{n}} \right) \approx 1 + \frac{s^2}{2n}$$
Proof: To see why, recall that the Taylor series of a function $f(u)$ around the point $u = 0$ has the form

$$f(u) = \sum_{k=0}^{\infty} \frac{u^k f^{(k)}(0)}{k!} = f(0) + uf^{(1)}(0) + u^2 \frac{f^{(2)}(0)}{2} + o(u^2),$$

where $o(u^2) \to 0$ faster than $u^2$ and the derivatives

$$f^{(k)}(0) = \frac{d^k}{du^k} f(u) \bigg|_{u=0}$$

are assumed to exist.

Combining the Taylor series expansion with the moment-generating property of the transform for a r.v. $Y$ yields

$$M_Y(s) = \sum_{k=0}^{\infty} s^k \frac{M_Y^{(k)}(0)}{k!} = \sum_{k=0}^{\infty} s^k \frac{\mathbb{E}(Y^k)}{k!} = 1 + s \mathbb{E}(Y) + \frac{s^2}{2} \mathbb{E}(Y^2) + o(s^2)$$

Now, replacing $s$ by $s/\sqrt{n}$, we obtain

$$M_Y \left( \frac{s}{\sqrt{n}} \right) \approx 1 + \frac{s^2}{2n},$$

as claimed.
• Hence
\[
\lim_{n \to \infty} M_{Z_n}(s) = \lim_{n \to \infty} \left(1 + \frac{s^2}{2n}\right)^n = e^{s^2/2}
\]
So asymptotically the transform of \( Z_n \) looks like that for a \( \mathcal{N}(0, 1) \) r.v.

• The CLT follows from this result and the following fact:
If the sequence of transforms \( M_{Z_n}(s) \) converges to a transform \( M_Z(s) \) of a r.v. \( Z \) whose cdf is continuous, then the sequence of cdfs \( F_{Z_n} \) converges to the cdf of \( Z \)

The proof of this fact is beyond the scope of this course

---

**Confidence Intervals Revisited**

• Let \( X_1, X_2, \ldots, X_n \) be i.i.d. with finite mean \( E(X) \) and variance \( \sigma_X^2 \) and \( S_n \) be the sample mean

• Given \( \epsilon, \delta > 0 \), how large should the number of samples \( n \) be such that
\[
P\{|S_n - E(X)| \leq \epsilon\} \geq 1 - \delta,
\]

• Let’s use the CLT to find an estimate of \( n \). Consider:
\[
P\{|S_n - E(X)| \leq \epsilon\} = P\left\{\left|\frac{1}{n} \sum_{i=1}^{n} (X_i - E(X))\right| \leq \epsilon\right\}
= P\left\{\left|\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - E(X)}{\sigma_X}\right)\right| \leq \frac{\epsilon \sqrt{n}}{\sigma_X}\right\}
\approx 1 - 2Q\left(\frac{\epsilon \sqrt{n}}{\sigma_X}\right)
\]
Example: For $\epsilon = 0.1\sigma_X$ and $\delta = .001$, we set $2Q(0.1\sqrt{n}) = .001$, which gives $0.1\sqrt{n} = 3.3$, or $n = 1089 (< < 10^5$, the bound using Chebyshev inequality)

Example: The Pollster’s Problem Revisited

Recall that $p$ is the fraction of the population that intend to vote for Snidely Whiplash against Dudley Doright. For the $i$th person polled

$$X_i = \begin{cases} 1 & \text{prefers Whiplash} \\ 0 & \text{otherwise} \end{cases}$$

$S_n$ is the sample average of the $X_i$s

The pollster wants to choose $n$ large enough to ensure that

$$P\{|S_n - p| \leq 0.01\} \geq 0.95$$

Before, we used Chebyshev inequality to obtain

$$P\{|S_n - p| \leq 0.01\} \geq 1 - \frac{\sigma_X^2}{n(0.01)^2}$$

and concluded that $n = 50,000$ suffices

Now let’s use the CLT to approximate $P\{|S_n - p| \leq .01\}$. Consider

$$P\{|S_n - p| \leq 0.01\} = P\left\{\left| \frac{1}{n} \sum_{i=1}^{n} (X_i - E(X)) \right| \leq 0.01 \right\}$$

$$= P\left\{ \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \frac{X_i - E(X)}{\sigma_X} \right) \right| \leq \frac{0.01\sqrt{n}}{\sigma_X} \right\}$$

$$= P\left\{ |Z_n| \leq \frac{0.01\sqrt{n}}{\sigma_X} \right\}$$

Since $\sigma_X \leq 1/2$,

$$P\{|S_n - p| \leq 0.01\} \geq P\{|Z_n| \leq 0.02\sqrt{n}\} \approx 1 - 2Q(0.02\sqrt{n})$$

Thus to have $P\{|S_n - p| \leq .01\} \geq 0.95$, we set $2Q(0.02\sqrt{n}) = 0.05$, which, using the $Q$-function table, gives $n \approx 9604$; a significantly better answer than the previous one
Random Processes

- A random process (also called stochastic process) \(\{X(t) : t \in T\}\) is an infinite collection of random variables, one for each value of time \(t \in T\) (or, in some cases distance).

- Random processes are used to model random experiments that evolve in time:
  - Received sequence/waveform at the output of a communication channel
  - Packet arrival times at a node in a communication network
  - Thermal noise in a resistor
  - Scores of an NBA team in consecutive games
  - Daily price of a stock
  - Winnings or losses of a gambler
  - Earth movement around a fault line
Questions Involving Random Processes

- Dependencies of the random variables of the process:
  - How do future received values depend on past received values?
  - How do future prices of a stock depend on its past values?
  - How well do past earth movements predict an earthquake?

- Long term averages:
  - What is the proportion of time a queue is empty?
  - What is the average noise power generated by a resistor?

- Extreme or boundary events:
  - What is the probability that a link in a communication network is congested?
  - What is the probability that the maximum power in a power distribution line is exceeded?
  - What is the probability that a gambler will lose all his capital?

Discrete vs. Continuous-Time Processes

- The random process \( \{X(t) : t \in T\} \) is said to be discrete-time if the index set \( T \) is countably infinite, e.g., \( \{1, 2, \ldots\} \) or \( \{\ldots, -2, -1, 0, +1, +2, \ldots\} \):
  - The process is simply an infinite sequence of r.v.s \( X_1, X_2, \ldots \)
  - An outcome of the process is simply a sequence of numbers

- The random process \( \{X(t) : t \in T\} \) is said to be continuous-time if the index set \( T \) is a continuous set, e.g., \((0, \infty)\) or \((\infty, \infty)\):
  - The outcomes are random waveforms or random occurrences in continuous time

- We only discuss discrete-time random processes:
  - IID processes
  - Bernoulli process and associated processes
  - Markov processes
  - Markov chains
IID Processes

• A process $X_1, X_2, \ldots$ is said to be *independent and identically distributed* (IID, or i.i.d.) if it consists of an infinite sequence of independent and identically distributed random variables

• Two important examples:
  - Bernoulli process: $X_1, X_2, \ldots$ are i.i.d. $\text{Bern}(p), 0 < p < 1$, r.v.s. Model for random phenomena with binary outcomes, such as:
    * Sequence of coin flips
    * Noise sequence in a binary symmetric channel
    * The occurrence of random events such as packets (1 corresponding to an event and 0 to a non-event) in discrete-time
    * Binary expansion of a random number between 0 and 1
  - Discrete-time white Gaussian noise (WGN) process: $X_1, X_2, \ldots$ are i.i.d. $\mathcal{N}(0, N)$ r.v.s. Model for:
    * Receiver noise in a communication system
    * Fluctuations in a stock price

• Useful properties of an IID process:
  - *Independence*: Since the r.v.s in an IID process are independent, any two events defined on sets of random variables with *non-overlapping* indices are independent
  - *Memorylessness*: The independence property implies that the IID process is memoryless in the sense that for any time $n$, the future $X_{n+1}, X_{n+2}, \ldots$ is independent of the past $X_1, X_2, \ldots, X_n$
  - *Fresh start*: Starting from any time $n$, the random process $X_n, X_{n+1}, \ldots$ behaves identically to the process $X_1, X_2, \ldots$, i.e., it is also an IID process with the same distribution. This property follows from the fact that the r.v.s are identically distributed (in addition to being independent)
The Bernoulli Process

- The Bernoulli process is an infinite sequence $X_1, X_2, \ldots$ of i.i.d. $\text{Bern}(p)$ r.v.s.
- The outcome from a Bernoulli process is an infinite sequence of 0s and 1s.
- A Bernoulli process is often used to model occurrences of random events; $X_n = 1$ if an event occurs at time $n$, and 0, otherwise.
- Three associated random processes of interest:
  - Binomial counting process: The number of events in the interval $[1, n]$.
  - Arrival time process: The time of event arrivals.
  - Interarrival time process: The time between consecutive event arrivals.
- We discuss these processes and their relationships.

Binomial Counting Process

- Consider a Bernoulli process $X_1, X_2, \ldots$ with parameter $p$.
- We are often interested in the number of events occurring in some time interval.
- For the time interval $[1, n]$, i.e., $i = 1, 2, \ldots, n$, we know that the number of occurrences
  \[ W_n = \left( \sum_{i=1}^{n} X_i \right) \sim \text{B}(n, p) \]
- The sequence of r.v.s $W_1, W_2, \ldots$ is referred to as a Binomial counting process.
- The Bernoulli process can be obtained from the Binomial counting process as:
  \[ X_n = W_n - W_{n-1}, \text{ for } n = 1, 2, \ldots, \]
  where $W_0 = 0$.
- Outcomes of a Binomial process are integer valued stair-case functions.
• Note that the Binomial counting process is not IID

• By the fresh-start property of the Bernoulli process, for any $n \geq 1$ and $k \geq 1$, the distribution of the number of events in the interval $[k + 1, n + k]$ is identical to that of $[1, n]$, i.e., $W_n$ and $(W_{k+n} - W_k)$ are identically distributed

---

**Example:** Packet arrivals at a node in a communication network can be modeled by a Bernoulli process with $p = 0.09$.

1. What is the probability that 3 packets arrive in the interval $[1, 20]$, 6 packets arrive in $[1, 40]$ and 12 packets arrive in $[1, 80]$?

2. The input queue at the node has a capacity of $10^3$ packets. A packet is dropped if the queue is full. What is the probability that one or more packets are dropped in a time interval of length $n = 10^4$?

**Solution:** Let $W_n$ be the number of packets arriving in interval $[1, n]$.

1. We want to find the following probability

$$P\{W_{20} = 3, W_{40} = 6, W_{80} = 12\}$$

which is equal to

$$P\{W_{20} = 3, W_{40} - W_{20} = 3, W_{80} - W_{40} = 6\}$$

By the independence property of the Bernoulli process this is equal to

$$P\{W_{20} = 3\}P\{W_{40} - W_{20} = 3\}P\{W_{80} - W_{40} = 6\}$$
Now, by the fresh start property of the Bernoulli process
\[
P\{W_{40} - W_{20} = 3\} = P\{W_{20} = 3\}, \quad \text{and} \quad P\{W_{80} - W_{40} = 6\} = P\{W_{40} = 6\}
\]

Thus
\[
P\{W_{20} = 3, W_{40} = 6, W_{80} = 12\} = (P\{W_{20} = 3\})^2 \times P\{W_{40} = 6\}
\]
Now, using the Poisson approximation of Binomial, we have
\[
P\{W_{20} = 3\} = \binom{20}{3}(0.09)^3(0.91)^{17} \approx \frac{(1.8)^3}{3!}e^{-1.8} = 0.1607
\]
\[
P\{W_{40} = 6\} = \binom{40}{6}(0.09)^6(0.91)^{34} \approx \frac{(3.6)^6}{6!}e^{-3.6} = 0.0826
\]
Thus
\[
P\{W_{20} = 3, W_{40} = 6, W_{80} = 12\} \approx (0.1607)^2 \times 0.0826 = 0.0021
\]

2. The probability that one or more packets are dropped in a time interval of length \(n = 10^4\) is
\[
P\{W_{10^4} > 10^3\} = \sum_{n=1001}^{10^4} \binom{10^4}{n}(0.09)^n(0.91)^{10^4-n}
\]
Difficult to compute, but we can use the CLT!
Since \(W_{10^4} = \sum_{i=1}^{10^4} X_i\) and \(E(X) = 0.09\) and \(\sigma^2_X = 0.09 \times 0.91 = 0.0819\), we have
\[
P\left\{ \sum_{i=1}^{10^4} X_i > 10^3 \right\} = P\left\{ \frac{1}{100} \sum_{i=1}^{10^4} \frac{(X_i - 0.09)}{\sqrt{0.0819}} > \frac{10^3 - 900}{100\sqrt{0.0819}} \right\}
\]
\[
= P\left\{ \frac{1}{100} \sum_{i=1}^{10^4} \frac{(X_i - 0.09)}{0.286} > 3.5 \right\}
\]
\[
\approx Q(3.5) = 2 \times 10^{-4}
\]
• Again consider a Bernoulli process $X_1, X_2, \ldots$ as a model for random arrivals of events.

• Let $Y_k$ be the time index of the $k$th arrival, or the $k$th arrival time, i.e., smallest $n$ such that $W_n = k$.

• Define the interarrival time process associated with the Bernoulli process as

$$T_1 = Y_1 \text{ and } T_k = Y_k - Y_{k-1}, \text{ for } k = 2, 3, \ldots$$

Thus the $k$th arrival time is given by: $Y_k = T_1 + T_2 + \ldots + T_k$

• Let’s find the pmf of $T_k$:

First, the pmf of $T_1$ is the same as the number of coin flips until a head (i.e, a 1) appears. We know that this is $\text{Geom}(p)$. Thus $T_1 \sim \text{Geom}(p)$

Now, having an event at time $T_1$, the future is a fresh starting Bernoulli process. Thus, the number of trials $T_2$ until the next event has the same pmf as $T_1$

Moreover, $T_1$ and $T_2$ are independent, since the trials from 1 to $T_1$ are independent of the trials from $T_1 + 1$ onward. Since $T_2$ is determined exclusively by what happens in these future trials, it’s independent of $T_1$

Continuing similarly, we conclude that $T_1, T_2, \ldots$ are i.i.d., i.e., the interarrival process is an IID $\text{Geom}(p)$ process.

• The interarrival process gives us an alternate definition of a Bernoulli process:

Start with an IID $\text{Geom}(p)$ process $T_1, T_2, \ldots$. Record the arrival of an event at time $T_1, T_1 + T_2, T_1 + T_2 + T_3, \ldots$
• Arrival time process: The sequence of r.v.s $Y_1, Y_2, \ldots$ is denoted by the arrival time process. From its relationship to the interarrival time process $Y_1 = T_1, Y_k = \sum_{i=1}^{k} T_i$, we can easily find the mean and variance of $Y_k$ for any $k$

$$E(Y_k) = E\left(\sum_{i=1}^{k} T_i\right) = \sum_{i=1}^{k} E(T_i) = k \times \frac{1}{p}$$

$$Var(Y_k) = Var\left(\sum_{i=1}^{k} T_i\right) = \sum_{i=1}^{k} Var(T_i) = k \times \frac{1 - p}{p^2}$$

Note that, $Y_1, Y_2, \ldots$ is not an IID process.

It is also not difficult to show that the pmf of $Y_k$ is

$$p_{Y_k}(n) = \binom{n-1}{k-1} p^k (1-p)^{n-k} \text{ for } n = k, k+1, k+2, \ldots,$$

which is called the Pascal pmf of order $k$.

---

• Example: In each minute of a basketball game, Alicia commits a foul independently with probability $p$ and no foul with probability $1 - p$. She stops playing if she commits her sixth foul or plays a total of 30 minutes. What is the pmf of Alicia's playing time?

Solution: We model the foul events as a Bernoulli process with parameter $p$.

Let $Z$ be the time Alicia plays. Then

$$Z = \min\{Y_6, 30\}$$

The pmf of $Y_6$ is

$$p_{Y_6}(n) = \binom{n-1}{5} p^6 (1-p)^{n-6}, \text{ for } n = 6, 7, \ldots$$

Thus the pmf of $Z$ is

$$p_Z(z) = \begin{cases} 
\binom{z-1}{5} p^6 (1-p)^{z-6}, & \text{for } z = 6, 7, \ldots, 29 \\
1 - \sum_{z=6}^{29} p_Z(z), & \text{for } z = 30 \\
0, & \text{otherwise} 
\end{cases}$$
Markov Processes

• A discrete-time random process $X_0, X_1, X_2, \ldots$, where the $X_n$s are discrete-valued r.v.s, is said to be a Markov process if for all $n \geq 0$ and all $(x_0, x_1, x_2, \ldots, x_n, x_{n+1})$

\[ P\{X_{n+1} = x_{n+1}|X_n = x_n, \ldots, X_0 = x_0\} = P\{X_{n+1} = x_{n+1}|X_n = x_n\}, \]

i.e., the past, $X_{n-1}, \ldots, X_0$, and the future, $X_{n+1}$, are conditionally independent given the present $X_n$

• A similar definition for continuous-valued Markov processes can be provided in terms of pdfs

• Examples:
  - Any IID process is Markov
  - The Binomial counting process is Markov

Markov Chains

• A discrete-time Markov process $X_0, X_1, X_2, \ldots$ is called a Markov chain if
  - For all $n \geq 0$, $X_n \in S$, where $S$ is a finite set called the state space.
    We often assume that $S \in \{1, 2, \ldots, m\}$
  - For $n \geq 0$ and $i, j \in S$

\[ P\{X_{n+1} = j|X_n = i\} = p_{ij}, \text{ independent of } n \]

So, a Markov chain is specified by a transition probability matrix

\[
\mathcal{P} = \begin{bmatrix}
  p_{11} & p_{12} & \cdots & p_{1m} \\
  p_{21} & p_{22} & \cdots & p_{2m} \\
  \vdots & \vdots & \ddots & \vdots \\
  p_{m1} & p_{m2} & \cdots & p_{mm}
\end{bmatrix}
\]

Clearly $\sum_{j=1}^{m} p_{ij} = 1$, for all $i$, i.e., the sum of any row is 1
• By the Markov property, for all \( n \geq 0 \) and all states

\[
P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} = P\{X_{n+1} = j | X_n = i\} = p_{ij}
\]

• Markov chains arise in many real world applications:
  ◦ Computer networks
  ◦ Computer system reliability
  ◦ Machine learning
  ◦ Pattern recognition
  ◦ Physics
  ◦ Biology
  ◦ Economics
  ◦ Linguistics

Examples

• Any IID process with discrete and finite-valued r.v.s is a Markov chain

• Binary Symmetric Markov Chain: Consider a sequence of coin flips, where each flip has probability of \( 1 - p \) of having the same outcome as the previous coin flip, regardless of all previous flips

The probability transition matrix (head is 1 and tail is 0) is

\[
P = \begin{bmatrix}
1 - p & p \\
p & 1 - p
\end{bmatrix}
\]

A Markov chain can be specified by a transition probability graph

Nodes are states, arcs are state transitions; \((i, j)\) from state \(i\) to state \(j\) (only draw transitions with \(p_{ij} > 0\))
Can construct this process from an IID process: Let $Z_1, Z_2, \ldots$ be a Bernoulli process with parameter $p$

The Binary symmetric Markov chain $X_n$ can be defined as

$$X_{n+1} = X_n + Z_n \mod 2,$$

for $n = 1, 2, \ldots$,

So, each transition corresponds to passing the r.v. $X_n$ through a binary symmetric channel with additive noise $Z_n \sim \text{Bern}(p)$

- Example: *Asymmetric binary Markov chain*

  State 0 = Machine is working, State 1 = Machine is broken down

  The probability transition matrix is

  $$P = \begin{bmatrix}
  0.8 & 0.2 \\
  0.6 & 0.4 
\end{bmatrix}$$

  and the transition probability graph is:

  ![Transition Probability Graph](image)

- Example: *Two spiders and a fly*

  A fly’s possible positions are represented by four states

  States 2, 3 : safely flying in the left or right half of a room
  State 1: A spider’s web on the left wall
  State 4: A spider’s web on the right wall

  The probability transition matrix is:

  $$P = \begin{bmatrix}
  1.0 & 0 & 0 & 0 \\
  0.3 & 0.4 & 0.3 & 0 \\
  0 & 0.3 & 0.4 & 0.3 \\
  0 & 0 & 0 & 1.0 
\end{bmatrix}$$

  The transition probability graph is:

  ![Transition Probability Graph](image)
Given a Markov chain model, we can compute the probability of a sequence of states given an initial state $X_0 = i_0$ using the chain rule as:

$$P\{X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n | X_0 = i_0\} = p_{i_0i_1}p_{i_1i_2}\cdots p_{i_{n-1}i_n}$$

Example: For the spider and fly example,

$$P\{X_1 = 2, X_2 = 2, X_3 = 3 | X_0 = 2\} = p_{22}p_{22}p_{23}$$

$$= 0.4 \times 0.4 \times 0.3$$

$$= 0.048$$

There are many other questions of interest, including:

- $n$-state transition probabilities: Beginning from some state $i$ what is the probability that in $n$ steps we end up in state $j$?

- Steady state probabilities: What is the expected fraction of time spent in state $i$ as $n \to \infty$?

### $n$-State Transition Probabilities

- Consider an $m$-state Markov Chain. Define the $n$-step transition probabilities as

$$r_{ij}(n) = P\{X_n = j | X_0 = i\} \text{ for } 1 \leq i, j \leq m$$

- The $n$-step transition probabilities can be computed using the *Chapman-Kolmogorov* recursive equation:

$$r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1)p_{kj} \text{ for } n > 1, \text{ and all } 1 \leq i, j \leq m,$$

starting with $r_{ij}(1) = p_{ij}$

This can be readily verified using the law of total probability
• We can view the $r_{ij}(n), 1 \leq i, j \leq m$, as the elements of a matrix $R(n)$, called the $n$-step transition probability matrix, then we can view the Kolmogorov-Chapman equations as a sequence of matrix multiplications:

\[
R(1) = \mathcal{P} \\
R(2) = R(1)\mathcal{P} = \mathcal{P}\mathcal{P} = \mathcal{P}^2 \\
R(3) = R(2)\mathcal{P} = \mathcal{P}^3 \\
\vdots \\
R(n) = R(n-1)\mathcal{P} = \mathcal{P}^n
\]

• Example: For the binary asymmetric Markov chain

\[
R(1) = \mathcal{P} = \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} \\
R(2) = \mathcal{P}^2 = \begin{bmatrix} 0.76 & 0.24 \\ 0.72 & 0.28 \end{bmatrix} \\
R(3) = \mathcal{P}^3 = \begin{bmatrix} 0.752 & 0.248 \\ 0.744 & 0.256 \end{bmatrix} \\
R(4) = \mathcal{P}^4 = \begin{bmatrix} 0.7504 & 0.2496 \\ 0.7488 & 0.2512 \end{bmatrix} \\
R(5) = \mathcal{P}^5 = \begin{bmatrix} 0.7501 & 0.2499 \\ 0.7498 & 0.2502 \end{bmatrix}
\]

In this example, each $r_{ij}$ seems to converge to a non-zero limit independent of the initial state, i.e., each state has a steady state probability of being occupied as $n \to \infty$. 
Example: Consider the spiders-and-fly example

\[
P = \begin{bmatrix}
1.0 & 0 & 0 & 0 \\
0.3 & 0.4 & 0.3 & 0 \\
0 & 0.3 & 0.4 & 0.3 \\
0 & 0 & 0 & 1.0 \\
\end{bmatrix}
\]

\[
P^2 = \begin{bmatrix}
1.0 & 0 & 0 & 0 \\
0.42 & 0.25 & 0.24 & 0.09 \\
0.09 & 0.24 & 0.25 & 0.42 \\
0 & 0 & 0 & 1.0 \\
\end{bmatrix}
\]

\[
P^3 = \begin{bmatrix}
1.0 & 0 & 0 & 0 \\
0.50 & 0.17 & 0.17 & 0.16 \\
0.16 & 0.17 & 0.17 & 0.50 \\
0 & 0 & 0 & 1.0 \\
\end{bmatrix}
\]

\[
P^4 = \begin{bmatrix}
1.0 & 0 & 0 & 0 \\
0.55 & 0.12 & 0.12 & 0.21 \\
0.21 & 0.12 & 0.12 & 0.55 \\
0 & 0 & 0 & 1.0 \\
\end{bmatrix}
\]

As \( n \to \infty \), we obtain

\[
P^\infty = \begin{bmatrix}
1.0 & 0 & 0 & 0 \\
2/3 & 0 & 0 & 1/3 \\
1/3 & 0 & 0 & 2/3 \\
0 & 0 & 0 & 1.0 \\
\end{bmatrix}
\]

Here \( r_{ij} \) converges, but the limit depends on the initial state and can be 0 for some states.

Note that states 1 and 4 corresponding to capturing the fly by one of the spiders are absorbing states, i.e., they are infinitely repeated once visited.

The probability of being in non-absorbing states 2 and 3 diminishes as time increases.
Classification of States

• As we have seen, various states of a Markov chain can have different characteristics

• We wish to classify the states by the long-term frequency with which they are visited

• Let $A(i)$ be the set of states that are accessible from state $i$ (may include $i$ itself), i.e., can be reached from $i$ in $n$ steps, for some $n$

• State $i$ is said to be recurrent if starting from $i$, any accessible state $j$ must be such that $i$ is accessible from $j$, i.e., $j \in A(i)$ iff $i \in A(j)$. Clearly, this implies that if $i$ is recurrent then it must be in $A(i)$

• A state is said to be transient if it is not recurrent

• Note that recurrence/transience is determined by the arcs (transitions with nonzero probability), not by actual values of probabilities

• Example: Classify the states of the following Markov chain

![Markov chain diagram]

• The set of accessible states $A(i)$ from some recurrent state $i$ is called a recurrent class

• Every state $k$ in a recurrent class $A(i)$ is recurrent and $A(k) = A(i)$

Proof: Suppose $i$ is recurrent and $k \in A(i)$. Then $k$ is accessible from $i$ and hence, since $i$ is recurrent, $k$ can access $i$ and hence, through $i$, any state in $A(i)$. Thus $A(i) \subseteq A(k)$

Since $i$ can access $k$ and hence any state in $A(k)$, $A(k) \subseteq A(i)$. Thus $A(i) = A(k)$. This argument also proves that any $k \in A(i)$ is recurrent
- Two recurrent classes are either identical or disjoint

- Summary:
  - A Markov chain can be decomposed into one or more recurrent classes plus possibly some transient states
  - A recurrent state is accessible from all states in its class, but it is *not* accessible from states in other recurrent classes
  - A transient state is not accessible from any recurrent state
  - At least one recurrent state must be accessible from a given transient state

- Example: Find the recurrent classes in the following Markov chains

![Markov Chain Diagram](image)

EE 178: Random Processes
Periodic Classes

- A recurrent class $A$ is called **periodic** if its states can be grouped into $d > 1$ disjoint subsets $S_1, S_2, \ldots, S_d$, $\bigcup_{i=1}^{d} S_i = A$, such that all transitions from one subset lead to the next subset.

![Diagram of periodic classes]

- Example: Consider a Markov chain with probability transition graph

![Diagram of example Markov chain]

Note that the recurrent class $\{1, 2\}$ is periodic and for $i = 1, 2$

$$r_{ii}(n) = \begin{cases} 1 & \text{if } n \text{ is even} \\ 0 & \text{otherwise} \end{cases}$$

- Note that if the class is periodic, $r_{ii}(n)$ never converges to a steady-state.
Steady State Probabilities

- **Steady-state convergence theorem**: If a Markov chain has only one recurrent class and it is not periodic, then $r_{ij}(n)$ tends to a steady-state $\pi_j$ independent of $i$, i.e.,
  $$\lim_{n \to \infty} r_{ij}(n) = \pi_j \text{ for all } i$$

- **Steady-state equations**: Taking the limit as $n \to \infty$ of the Chapman-Kolmogorov equations
  $$r_{ij}(n+1) = \sum_{k=1}^{m} r_{ik}(n)p_{kj} \text{ for } 1 \leq i, j \leq m,$$
  we obtain the set of linear equations, called the *balance equations*:
  $$\pi_j = \sum_{k=1}^{m} \pi_k p_{kj} \text{ for } j = 1, 2, \ldots, m$$

  The balance equations together with the *normalization* equation
  $$\sum_{j=1}^{m} \pi_j = 1,$$

  uniquely determine the steady state probabilities $\pi_1, \pi_2, \ldots, \pi_m$

- The balance equations can be expressed in a matrix form as:
  $$\Pi P = \Pi, \text{ where } \Pi = [\pi_1 \pi_2 \ldots \pi_m]$$

- In general there are $m - 1$ linearly independent balance equations

- The steady state probabilities form a probability distribution over the state space, called the *stationary distribution* of the chain. If we set $P\{X_0 = j\} = \pi_j$ for $j = 1, 2, \ldots, m$, we have $P\{X_n = j\} = \pi_j$ for all $n \geq 1$ and $j = 1, 2, \ldots, m$

- **Example: Binary Symmetric Markov Chain**
  $$P = \begin{bmatrix} 1 - p & p \\ p & 1 - p \end{bmatrix}$$

  Find the steady-state probabilities

  Solution: We need to solve the balance and normalization equations
  $$\pi_1 = p_{11}\pi_1 + p_{21}\pi_2 = (1 - p)\pi_1 + p\pi_2$$
  $$\pi_2 = p_{12}\pi_1 + p_{22}\pi_2 = p\pi_1 + (1 - p)\pi_2$$
  $$1 = \pi_1 + \pi_2$$
Note that the first two equations are linearly dependent. Both yield \( \pi_1 = \pi_2 \).
Substituting in the last equation, we obtain \( \pi_1 = \pi_2 = 1/2 \)

- **Example:** *Asymmetric binary Markov chain*

\[
P = \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix}
\]

The steady state equations are:

\[
\pi_1 = p_{11}\pi_1 + p_{21}\pi_2 = 0.8\pi_1 + 0.6\pi_2, \text{ and } 1 = \pi_1 + \pi_2
\]

Solving the two equations yields \( \pi_1 = 3/4 \) and \( \pi_2 = 1/4 \)

- **Example:** Consider the Markov chain defined by the following probability transition graph

```
1 2 3 4
0.9 0.1 0.5 0.5 0.4
0.2 0.8
```

- **Note:** The solution of the steady state equations yields \( \pi_j = 0 \) if a state is transient, and \( \pi_j > 0 \) if a state is recurrent
Long Term Frequency Interpretations

• Let \( v_{ij}(n) \) be the # of times state \( j \) is visited beginning from state \( i \) in \( n \) steps

• For a Markov chain with a single aperiodic recurrent class, can show that

\[
\lim_{n \to \infty} \frac{v_{ij}(n)}{n} = \pi_j
\]

Since this result doesn’t depend on the starting state \( i \), \( \pi_j \) can be interpreted as the long-term frequency of visiting state \( j \)

• Since each time state \( j \) is visited, there is a probability \( p_{jk} \) that the next transition is to state \( k \), \( \pi_j p_{jk} \) can be interpreted as the long-term frequency of transitions from \( j \) to \( k \)

• These frequency interpretations allow for a simple interpretation of the balance equations, that is, the long-term frequency \( \pi_j \) is the sum of the long-term frequencies \( \pi_k p_{kj} \) of transitions that lead to \( j \)

\[
\pi_j = \sum_{k=1}^{m} \pi_k p_{kj}
\]

Another interpretation of the balance equations: Rewrite the LHS of the balance equation as

\[
\pi_j = \pi_j \sum_{k=1}^{m} p_{jk} = \sum_{k=1}^{m} \pi_j p_{jk} = \pi_j p_{jj} + \sum_{k=1, k \neq j}^{m} \pi_j p_{jk}
\]

The RHS can be written as

\[
\sum_{k=1}^{m} \pi_k p_{kj} = \pi_j p_{jj} + \sum_{k=1, k \neq j}^{m} \pi_k p_{kj}
\]

Subtracting the \( p_{jj} \pi_j \) from both sides yields

\[
\sum_{k=1, k \neq j}^{m} \pi_k p_{kj} = \sum_{k=1, k \neq j}^{m} \pi_j p_{jk}, \ j = 1, 2, \ldots, m
\]
The long-term frequency of transitions into $j$ is equal to the long-term frequency of transitions out of $j$

$$\pi_1 p_{1j} \quad \pi_j p_{j1}$$

$$\pi_2 p_{2j} \quad \pi_j p_{j2}$$

$$\vdots \quad \vdots$$

$$\pi_m p_{mj} \quad \pi_j p_{jm}$$

This interpretation is similar to Kirkoff’s current law. In general, if we partition a chain (with single aperiodic recurrent class) into two sets of states, the long-term frequency of transitions from the first set to the second is equal to the long-term frequency of transitions from the second to the first.

---

**Birth-Death Processes**

- A *birth-death* process is a Markov chain in which the states are linearly arranged and transitions can only occur to a neighboring state, or else leave the state unchanged.

- For a birth-death Markov chain use the notation:

  \[
  b_i = P\{X_{n+1} = i+1 | X_n = i\}, \quad \text{birth probability at state } i \\
  d_i = P\{X_{n+1} = i-1 | X_n = i\}, \quad \text{death probability at state } i \\
  \]

\[
\begin{align*}
1 - b_0 & \quad 1 - b_1 - d_1 \\
0 & \quad b_0 \quad b_1 \quad \ddots \\
& \quad d_1 \quad d_2 \\
1 & \quad \vdots \\
1 - b_{m-1} - d_{m-1} & \quad 1 - d_m \\
& \quad b_{m-2} \quad b_{m-1} \\
& \quad d_{m-1} \quad d_m \\
& \quad m-1 \quad m
\end{align*}
\]
For a birth-death process the balance equations can be greatly simplified: Cut the chain between states \( i - 1 \) and \( i \). The long-term frequency of transitions from right to left must be equal to the long-term frequency of transitions from left to right, thus:

\[
\pi_i d_i = \pi_{i-1} b_{i-1}, \text{ or } \pi_i = \pi_{i-1} \frac{b_{i-1}}{d_i}
\]

By recursive substitution, we obtain

\[
\pi_i = \pi_0 b_0 b_1 \ldots b_{i-1} \frac{d_1 d_2 \ldots d_i}{d_1 d_2 \ldots d_i}, \quad i = 1, 2, \ldots, m
\]

To obtain the steady state probabilities we use these equations together with the normalization equation

\[
1 = \sum_{j=0}^{m} \pi_j
\]

**Examples**

- Queuing: Packets arrive at a communication node with buffer size \( m \) packets. Time is discretized in small periods. At each period:
  - If the buffer has less than \( m \) packets, the probability of 1 packet added to it is \( b \), and if it has \( m \) packets, the probability of adding another packet is 0
  - If there is at least 1 packet in the buffer, the probability of 1 packet leaving it is \( d > b \), and if it has 0 packets, this probability is 0
  - If the number of packets in the buffer is from 1 to \( m - 1 \), the probability of no change in the state of the buffer is \( 1 - b - d \). If the buffer has no packets, the probability of no change in the state is \( 1 - b \), and if there are \( m \) packets in the buffer, this probability is \( 1 - d \)

We wish to find the long-term frequency of having \( i \) packets in the queue
We introduce a birth-death Markov chain with states 0, 1, \ldots, m, corresponding to the number of packets in the buffer

The local balance equations are

\[ \pi_i d = \pi_{i-1} b, \quad i = 1, \ldots, m \]

Define \( \rho = b/d < 1 \), then \( \pi_i = \rho \pi_{i-1} \), which leads to

\[ \pi_i = \rho^i \pi_0 \]

Using the normalizing equation: \( \sum_{i=0}^{m} \pi_i = 1 \), we obtain

\[ \pi_0 (1 + \rho + \rho^2 + \cdots + \rho^m) = 1 \]

Hence for \( i = 1, \ldots, m \)

\[ \pi_i = \frac{\rho^i}{1 + \rho + \rho^2 + \cdots + \rho^m} \]

Using the geometric progression formula, we obtain

\[ \pi_i = \rho^i \frac{1 - \rho}{1 - \rho^{m+1}} \]

Since \( \rho < 1 \), \( \pi_i \rightarrow \rho^i (1 - \rho) \) as \( m \rightarrow \infty \), i.e., \( \{\pi_i\} \) converges to Geometric pmf

- **The Ehrenfest model**: This is a Markov chain arising in statistical physics. It models the diffusion through a membrane between two containers. Assume that the two containers have a total of 2\( a \) molecules. At each step a molecule is selected at random and moved to the other container (so a molecule diffuses at random through the membrane). Let \( Y_n \) be the number of molecules in container 1 at time \( n \) and \( X_n = Y_n - a \). Then \( X_n \) is a birth-death Markov chain with 2\( a + 1 \) states; \( i = -a, -a + 1, \ldots, -1, 0, 1, 2, \ldots, a \) and probability transitions

\[ p_{ij} = \begin{cases} b_i = (a - i)/2a, & \text{if } j = i + 1 \\ d_i = (a + i)/2a, & \text{if } j = i - 1 \\ 0, & \text{otherwise} \end{cases} \]
The steady state probabilities are given by:

\[ \pi_i = \pi_{-a} \frac{b_{-a}b_{-a+1}\cdots b_{i-1}}{d_{-a+1}d_{-a+2}\cdots d_i}, \quad i = -a, -a+1, \ldots, -1, 0, 1, 2, \ldots, a \]

\[ = \pi_{-a} \frac{2a(2a - 1)\cdots (a - i + 1)}{1 \times 2 \times \ldots \times (a + i)} \]

\[ = \pi_{-a} \frac{2a!}{(a + i)!(2a - (a + i))!} = \left( \frac{2a}{a + i} \right) \pi_{-a} \]

Now the normalization equation gives

\[ \sum_{i=-a}^{a} \left( \frac{2a}{a + i} \right) \pi_{-a} = 1 \]

Thus, \( \pi_{-a} = 2^{-2a} \)

Substituting, we obtain

\[ \pi_i = \left( \frac{2a}{a + i} \right) 2^{-2a} \]