Lectures on Probability and Mathematical Statistics

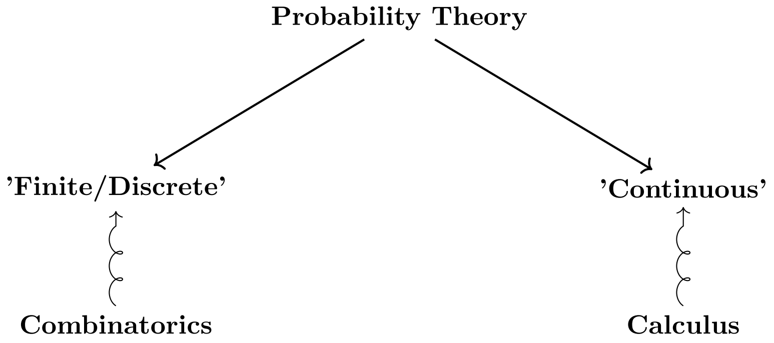
Contents

Lecture 0
Introduction to Probability

Today, we are setting sail into the intriguing realm of probability, a branch of mathematics that helps us navigate uncertainty and randomness.

Introduction to Probability Theory

Probability theory can be broadly categorized into two main branches: 'Finite/Discrete' and 'Continuous'. These branches are distinguished by the mathematical tools they employ, with combinatorics playing a key role in the former and calculus in the latter:



Finite/Discrete Probability

In finite or discrete probability, we deal with a finite number of distinct outcomes. The main tool here is combinatorics, which involves counting principles.

Example: Flipping a Coin

Consider the simple act of flipping a coin. If we were to ask, 'What is the probability of getting heads?' without any further information, we would find ourselves in a bit of a conundrum. Why? Because without knowing how likely it is for the coin to land on heads versus tails, we can't assign a meaningful probability.

To make sense of this, think about the scenario in which the coin lands near a wall, wobbles, but never fully flips onto a side. In this case, it wouldn't be accurate to say that the outcome is either heads or tails; rather, it's undetermined.

Now, let's make some additional assumptions. Suppose we have a fair coin, one that's perfectly balanced, and we flip it once. In this case, we have two possible outcomes: heads or tails. Since the coin is fair, we assume that the likelihood of getting heads is the same as getting tails, each with a probability of 0.5 or 50%.

What if we decide to flip the coin twice? Now, the possible outcomes expand. We could get heads on both flips, tails on both flips, or a combination of heads and tails. There are four possible outcomes: $\{$(H, H), (H, T), (T, H), (T, T)$\}$. Assuming the coin flips are independent (the outcome of one flip doesn't affect the outcome of the next), each of these outcomes has a probability of 0.25 or 25%.

This illustrates how probabilities come into play in simple situations like coin flips. By making appropriate assumptions and considering the possible outcomes, we can assign meaningful probabilities to events.

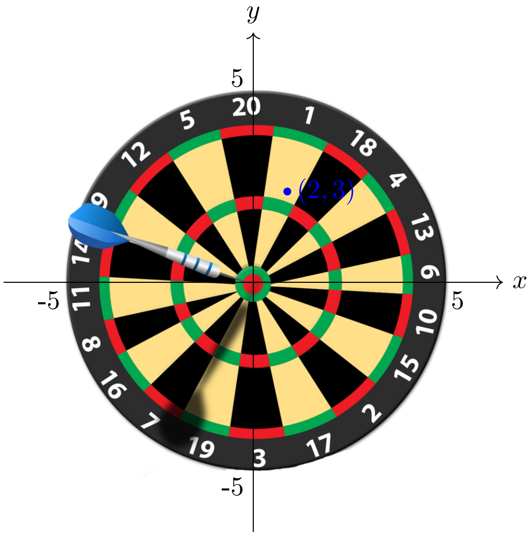
Continuous Probability

In continuous probability, we deal with outcomes from a continuous set, such as real numbers within a certain range. The main tool here is calculus.

Example: Throwing a Dart

When we throw a dart at a circular target, the set of possible outcomes encompasses all points within the circle (assuming we never miss the target!).

Question. What is the probability of hitting the point $(2, 3)$?



This question, as stated, doesn't quite make sense yet. To assign probabilities, we need a distribution, which essentially tells us the likelihood of hitting different points. For simplicity, we can assume a 'uniform' distribution. In pedestrian terms, this means that each point within the circle has an equal chance of being hit.

Here's a fascinating paradox to consider: when we throw a dart, we do actually hit some point on the board. However, under this uniform distribution assumption, the probability of hitting any specific point, like $(2, 3)$, is zero. This is because there are infinitely many points in the circle, and the 'target' at each specific point is effectively a single point with zero area.

This concept aligns with the intuition that hitting an exact point on the target, out of countless possible points, is an extremely rare event. It's a paradox that challenges our everyday understanding of probabilities!

Interestingly, the probability of hitting a region, like the bullseye, can be calculated using calculus. It's simply the ratio of the area of the bullseye to the total area of the target.

Lecture 1
Set Theory Unraveled: From Elements to Unions

As we explore the subject of probability,, we find ourselves faced with outcomes and particular classes of events. Now, to accurately depict and scrutinize these situations, we need the perfect language. And that's where set theory steps in, setting things up the right way for us. It offers a robust framework to define, categorize, and work with these outcomes and events.

What is a Set?

A **set** refers to a collection of objects identified by particular properties. It is represented by $\mathcal{S} = \{\text{objects} \,|\, \text{properties}\}$. The **elements** of a set are the objects it contains. We use the notation $x \in \mathcal{S}$ to denote that x is an element of set $\mathcal{S}$. For instance, $2 \in \{1, 2, 3, 4, 5\}$.

Example. 1. The set of natural numbers: $\mathbb{N} = \{0, 1, 2, 3, 4, \ldots\}$.

2. The set of even natural numbers: $\mathbb{E} = \{x \in \mathbb{N} \,|\, x \text{ is an even integer}\}$.

3. The set of days in a week: $\mathcal{D} = \{\text{Sunday}, \text{Monday}, \text{Tuesday}, \text{Wednesday}, \text{Thursday}, \text{Friday}, \text{Saturday}\}$.

4. The set of days in a week when we have a class:

$$\mathcal{DC} = \{x \in \mathcal{D} \,|\, \text{we have a class}\} = \{\text{Monday}, \text{Tuesday}, \text{Wednesday}, \text{Thursday}\}.$$

Subsets and Supersets

A **subset** $A \subseteq B$ means that every element in A is also in B. If $A \subset B$, then A is a subset of B but not equal to B. Similarly, $A \supseteq B$ indicates that every element in B is also in A, and $A \supset B$ means A is a **superset** of B but not equal to B.

Let's consider a few real-life examples:

Fruits and Apples. Set A can represent the set of all fruits in a grocery store. Set B can represent the set of all apples in that store. Since every apple is a fruit, we have $B \subseteq A$.

Math Courses and Calculus. Set A could be the set of all math courses offered in a university, and set B could be the set of all calculus courses. In this case, $B \subseteq A$ because all calculus courses are math courses, but not all math courses are necessarily calculus courses.

Animals and Dogs. Set A represents all animals on a farm, and set B represents all the dogs on that farm. Here, $B \subseteq A$ since every dog is an animal.

Operations on Sets

We introduce several basic operations on sets—such as union, intersection, and difference—that allow us to build new sets from existing ones and to describe relationships between them.

Definition. 1. The **union** of sets A and B, denoted $A \cup B$, is the set containing all elements that belong to A, or to B, or to both.

2. The **intersection** of sets A and B, denoted $A \cap B$, is the set containing all elements that belong to both A and B.

3. The **difference** of sets A and B, denoted $A \setminus B$, is the set of all elements that belong to A but not to B. It is sometimes read as "A minus B".

4. The **universal set**, denoted by $\mathcal{U}$, is the set that contains all elements under consideration for a particular discussion or problem. The **complement** of set A with respect to the universal set $\mathcal{U}$, denoted A^c, contains all elements that are in $\mathcal{U}$ but not in A.

5. Two sets A and B are called **disjoint** if their intersection is the empty set, i.e., $A \cap B = \varnothing$.

Example. 1. Consider the sets $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5, 6\}$.

Union: $A \cup B = \{1, 2, 3, 4, 5, 6\}$.

Intersection: $A \cap B = \{3, 4\}$.

Difference: $A \setminus B = \{1, 2\}$.

Complement: Assuming the universal set contains elements from 1 to 10, $A^c = \{1, 2, 5, 6, 7, 8, 9, 10\}$.

2. Let's consider sets $C = \{2, 4, 6, 8\}$ and $D = \{3, 6, 9\}$.

Union: $C \cup D = \{2, 3, 4, 6, 8, 9\}$.

Intersection: $C \cap D = \{6\}$.

Complement: Assuming the universal set contains numbers from 1 to 10, $D^c = \{1, 2, 4, 5, 7, 8, 10\}$.

3. Lastly, consider sets $X = \{1, 2, 3, 4\}$ and $Y = \{3, 4, 5\}$.

   Union: $X \cup Y = \{1, 2, 3, 4, 5\}$.
   Intersection: $X \cap Y = \{3, 4\}$.
   Complement: Assuming the universal set contains numbers from 1 to 5, $X^c = \{5\}$.

Unions and Intersections of Many Sets

Unions and intersections can involve not just a couple, but many sets. The union of multiple sets consists of elements that belong to at least one set from the collection under consideration, while the intersection consists of elements that are common to all sets in the collection. To deepen our understanding, let's explore some compelling examples involving countably infinite collections of sets.

Example. 1. Consider the sequence of sets $A_1, A_2, A_3, \ldots$, where for each positive integer n, we define

$$A_n = \{n, n+1, n+2, \ldots\}.$$

Each set A_n contains all integers greater than or equal to n.

- The sets are nested in a decreasing fashion:

  $$A_1 \supset A_2 \supset A_3 \supset \cdots,$$

  meaning each subsequent set is a proper subset of the previous one. The union of all the sets A_n collects all elements that appear in at least one of them. Since A_1 already contains all positive integers and every other set is contained in A_1, we have:

  $$\bigcup_{n=1}^{\infty} A_n = A_1 = \{1, 2, 3, \ldots\}.$$

- The intersection of all sets A_n includes only those elements that appear in every set. But no integer appears in all A_n, since for any fixed $m \in \mathbb{N}$, we can choose $n > m$, and $m \notin A_n$, so

  $$\bigcap_{n=1}^{\infty} A_n = \varnothing$$

  is the empty set.

2. Now consider the collection of open intervals

$$C_n = \left(-\frac{1}{n}, \frac{1}{n}\right) \quad \text{for } n \in \mathbb{N}.$$

Each C_n is a symmetric interval around 0 that gets narrower as n increases.

- Again, the sets are nested:

  $$C_1 \supset C_2 \supset C_3 \supset \cdots.$$

  The union of all C_n is

  $$\bigcup_{n=1}^{\infty} C_n = (-1, 1),$$

  because $C_1 = (-1, 1)$ already contains all the other sets in the collection.

- The intersection, however, is quite different. Any real number $x \neq 0$ lies outside C_n for sufficiently large n, since the endpoints $\pm\frac{1}{n}$ get arbitrarily close to 0. The only point that remains in every C_n is 0, therefore

  $$\bigcap_{n=1}^{\infty} C_n = \{0\}.$$

3. Finally, consider the collection of open intervals:

$$D_n = (-n, n), \quad \text{for } n \in \mathbb{N}.$$

Each interval grows wider with increasing n, eventually covering larger portion of the real line.

- In this case, the sets are nested in increasing way:

  $$D_1 \subset D_2 \subset D_3 \subset \cdots .$$

  The intersection of an increasing sequence of sets is just the smallest one:

  $$\bigcap_{n=1}^{\infty} D_n = D_1 = (-1, 1).$$

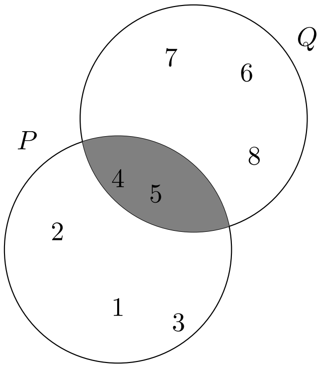
- Since $(-n, n)$ increases without bound and eventually covers every real number, any real number $x \in \mathbb{R}$ will be in some interval D_n for sufficiently large n, the union of all these expanding intervals is

  $$\bigcup_{n=1}^{\infty} D_n = \mathbb{R}.$$

Venn Diagrams

Venn diagrams provide a visual representation of sets and their relationships. For example, if $A \subseteq B$, we can draw a Venn diagram showing A inside B.

Example. Let's use Venn diagrams to illustrate set relationships.
Consider sets $P = \{1, 2, 3, 4, 5\}$ and $Q = \{4, 5, 6, 7, 8\}$.



In this Venn diagram, P is represented by the left circle and Q by the right circle. The overlapping region contains the elements that are in both P and Q, which are $\{4, 5\}$.

Cardinality and Power Set

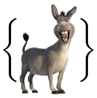
Definition. 1. The **cardinality** of a set A, denoted by $|A|$, represents the number of elements it contains.

2. The power set of a set A, denoted $\mathcal{P}(A)$, is the set of all subsets of A, including the empty set, $\varnothing$, and A itself.

Example. 1. Let's consider a set consisting of two *friendly* characters from a well-known animated film, Shrek and Donkey: $\mathcal{S} = \Bigg\{$, $\Bigg\}$.

Cardinality: The set $\mathcal{S}$ contains two distinct elements, so its cardinality is $|\mathcal{S}| = 2$.

Power Set: The power set $\mathcal{P}(\mathcal{S})$ is the set of all subsets of $\mathcal{S}$:

$\mathcal{P}(\mathcal{S}) = \Big\{\varnothing, \Big\{$ $\Big\}, \Big\{$  $\Big\}, \Big\{$ $,$ $\Big\}\Big\}.$

2. Consider the set $B = \{1, 2, 3\}$.

   Cardinality: The set B contains 3 elements, so the cardinality of B is $|B| = 3$.

   Power Set: The power set of B is

$$\mathcal{P}(B) = \{\varnothing, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}.$$

Question. Let A be a finite set with n elements. How many elements does the power set $\mathcal{P}(A)$ contain?

The elements of the power set $\mathcal{P}(A)$ are all the subsets of the set A. To count how many subsets there are, we observe that for each element $a \in A$, when forming a subset of A, we have exactly two choices:

- include a in the subset, or
- exclude a from the subset.

Since there are n elements in the set A, and each element has 2 independent choices (include or exclude), the total number of possible subsets is $|\mathcal{P}(A)| = 2^n$.

De Morgan's Laws

One of the most important identities in set theory involves how complements interact with unions and intersections. These identities, known as **De Morgan's Laws**, describe how the complement of a union becomes the intersection of complements, and how the complement of an intersection becomes the union of complements.

Definition. Let A and B be two subsets of some universal set U. Then the following identities hold:

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

Example. Let us understand De Morgan's law with the help of a simple example. Let the universal set $U = \{7, 8, 9, 10, 11, 12, 13\}$. The two subsets are given by $A = \{11, 12, 13\}$ and $B = \{7, 8\}$.

1. First, let's look at the complement of the union of A and B. We have

$$A \cup B = \{7, 8, 11, 12, 13\},$$
$$A^c = \{7, 8, 9, 10\},$$
$$B^c = \{9, 10, 11, 12, 13\}.$$

   Notice that the complement of $A \cup B$ is

$$(A \cup B)^c = \{9, 10\}.$$

   This shows that the complement of the union is exactly the intersection of the complements, just as De Morgan's Law says: $(A \cup B)^c = A^c \cap B^c$.

2. Next, consider the complement of the intersection of A and B:

$$A \cap B = \varnothing,$$
$$(A \cap B)^c = U = \{7, 8, 9, 10, 11, 12, 13\}.$$

   Observe that the union of complements is

$$A^c \cup B^c = \{7, 8, 9, 10, 11, 12, 13\},$$

   confirming the second De Morgan's Law: $(A \cap B)^c = A^c \cup B^c$.

Paradoxes (Optional Fun Facts)

Liar's Paradox

In the novel 'Don Quixote' written by Miguel de Cervantes, there's an intriguing scenario that includes a paradox from set theory known as the 'liar's paradox'. The book is highly recommended for its captivating tales and thought-provoking moments.

On the second day of his governing, Sancho eats a meager breakfast and goes into the courtroom. A man comes in and begins to describe a dilemma. A river cuts a lord's estate in two parts, and a bridge crosses over it. The owner of the river decreed that every person that wants to cross the bridge must state his purpose to several judges; if he tells the truth, he can cross, but if he lies, he must be hung. One man told the judges that his purpose is to be hung. If the judges allow him to cross, then his statement will have been a lie, and he should have been hung; if they hang him, then he told the truth, and he should have been allowed to cross. Sancho responds that this man deserves to live as much as he deserves to die, so it's better to be merciful and let him live. Everyone is satisfied with the decision.

The scenario presents a classic example of a self-referential paradox, often referred to as the "liar's paradox." The contradiction arises from the man's statement itself:

The man states, 'My purpose is to be hung'.

If the judges allow him to cross, then his statement is false, because he said he intended to be hung but was not. This would mean he should have been hung according to the rules.

On the other hand, if the judges hang him, then his statement is true, because he said he intended to be hung and he was. However, this also leads to a contradiction, because if he was telling the truth, then he should have been allowed to cross according to the rules.

So, no matter what the judges decide, they run into a logical contradiction. This creates a perplexing situation where it seems impossible to make a decision based on the rules established.

Sancho's response, that the man deserves to live as much as he deserves to die, is a recognition of this inherent contradiction. By showing mercy and allowing the man to live, Sancho essentially acknowledges that the situation is beyond a straightforward application of the rules.

This scenario illustrates the complexities and philosophical conundrums that can arise from self-referential statements and logical paradoxes, making it a thought-provoking moment in the novel.

Anticipated Movie Premiere Paradox

Imagine there's a highly anticipated movie premiere scheduled for next week. Fans have been eagerly waiting for this film for months, and the excitement is palpable. The organizers of the premiere have decided to add an element of surprise. They promised the tickets will go on sale on Wednesday, starting at an exact hour, but this hour will be impossible to be determined in advance.

Question. Given the information provided, when can the ticket sale start on Wednesday? Provide your analysis.

Exercises

Exercise 1. Consider the sets $X = \{a, b, c, d\}$ and $Y = \{c, d, e, f\}$. Find:

(a) $X \cup Y$

(b) $X \cap Y$

(c) X^c

Exercise 2. Let $A = \{1, 2, 3, 4, 5\}$ and $B = \{3, 4, 5, 6, 7\}$. Determine if A and B are disjoint sets.

Exercise 3. For sets $C = \{2, 4, 6, 8\}$ and $D = \{3, 6, 9\}$, find:

(a) $C \cup D$

(b) $C \cap D$

(c) C^c

(d) $C^c \cap D$

Exercise 4. Create a Venn diagram for sets $M = \{1, 2, 3, 4, 5\}$ and $N = \{4, 5, 6, 7, 8\}$. Shade the regions to represent $M \cup N$ and $M \cap N$.

Exercise 5. Consider sets $P = \{2, 4, 6, 8\}$, $Q = \{3, 6, 9\}$, and $R = \{1, 3, 5, 7\}$. Determine if $P \subseteq R$ and if $Q \cap R = \varnothing$.

Exercise 6. Let $E = \{a, b, c, d\}$ and $F = \{c, d, e, f\}$. Find the complement of the intersection of E and F, i.e., $(E \cap F)^c$.

Exercise 7. Use De Morgan's Laws to simplify the expression $(A \cap B)^c \cup (C \cap D)^c$, where $A = \{1, 2, 3\}$, $B = \{3, 4, 5\}$, $C = \{4, 5, 6\}$, and $D = \{6, 7, 8\}$.

Exercise 8. Suppose you have a sequence of time intervals representing the operating hours of different stores:

- Store A operates from 9:00 AM to 5:00 PM.
- Store B operates from 10:00 AM to 6:00 PM.
- Store C operates from 11:00 AM to 7:00 PM.
- Store D operates from 9:00 AM to 3:00 PM.

Define intervals S_A, S_B, S_C, and S_D to represent the operating hours of each store.

(a) Find the time interval during which at least one of the stores is open. Describe this interval in terms of hours. Then write the corresponding set in set theory notation (using S_A, S_B, S_C, and S_D with $\cup$ and $\cap$ operations).

(b) Identify the time period when all four stores are simultaneously open. Then write the corresponding set in set theory notation (using S_A, S_B, S_C, and S_D with $\cup$ and $\cap$ operations)..

Exercise 9. Consider the intervals $I_n = \left(\frac{1}{n+1}, \frac{1}{n}\right]$ for each positive integer n.

(*a*) Find the union $\bigcup_{n=1}^{\infty} I_n$.

(*b*) Find the intersection $\bigcap_{n=1}^{\infty} I_n$.

Lecture 2
Sample Spaces and Events

To begin, let's provide a more precise definition of the concept of a sample space, building upon our familiarity with set theory from the previous lecture.

Definition. A **sample space** is the set of all possible outcomes of an experiment.

Consider the following examples:

Rolling a six-sided die
Sample Space: {1, 2, 3, 4, 5, 6}

Drawing Cards from a Deck
Sample Space: {Ace of Spades, Two of Hearts, Jack of Diamonds, . . . }

Weather Forecasting
Sample Space: {Sunny, Cloudy, Rainy, Snowy}

Now, onto events.

Definition. An **event** is a subset of the sample space.

Let's take a look at some events related to our examples.

- For the Die Roll.
  - Event A: Rolling an even number = {2, 4, 6}
  - Event B: Rolling a number greater than 4 = {5, 6}
- For Drawing Cards.
  - Event C: Drawing a red card = $\{\heartsuit, \diamondsuit\}$
  - Event D: Drawing a face card = {Jack, Queen, King}

  Here's a bit more detail:

  - **Hearts:** This suit is symbolized by red heart shapes, $\heartsuit$. It consists of thirteen cards, ranging from Ace to King. These include numbered cards (2 through 10), face cards (Jack, Queen, King), and the Ace.
  - **Diamonds:** Similarly, this suit is represented by red diamond shapes, $\diamondsuit$. It also contains thirteen cards, comprising the same range as Hearts (Ace to King).

  So, when we talk about the event 'Drawing a red card', we're essentially referring to the combined set of Hearts and Diamonds. In other words, if you draw a card and it's either a Heart or a Diamond, it falls under the event 'Drawing a red card'.
- For Weather Forecasting:
  - Event E: Predicting a rainy day = {Rainy}
  - Event F: Expecting a non-sunny day = {Cloudy, Rainy, Snowy}

Probabilities as Functions

Now, probabilities come into play. These are functions that assign a likelihood to each event. They provide us with a numerical representation of how probable an outcome is. This is a crucial tool that aids us in making informed decisions, whether it's in games of chance, weather forecasting, or many other real-world applications.

So, as we venture further into probability theory, keep in mind that it's about understanding all potential outcomes and assigning values that signify their likelihood. This is the key to unlocking the power of probability.

We start with an exploration of the key characteristics that define probability functions.

Definition. A **probability function** on a sample space Ω is a function $P : \mathcal{P}(\Omega) \to [0, 1]$ that satisfies the following conditions:

1. **Normalization**: $P(\Omega) = 1$. This ensures that the total probability of all possible outcomes is 1.
2. **Additivity**: For any pair of disjoint sets A and B, $P(A \cup B) = P(A) + P(B)$. This is an extension of the idea that the probability of either of two mutually exclusive events occurring is the sum of their individual probabilities.

Let's explore some examples of probability functions.

Example: Fair Coin Toss

Consider a fair coin toss. In this simple experiment, the sample space Ω consists of two possible outcomes: heads (H) or tails (T). Mathematically, we represent this sample space as $\Omega = \{H, T\}$, where H represents heads and T represents tails.

We define $P(A)$ for any subset $A \subseteq \Omega$ as the ratio of the number of outcomes in A to the total number of outcomes in Ω. Mathematically, it is given by:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{2}$$

In particular, $P(\{H\}) = P(\{T\}) = \frac{1}{2}$.

This probability function satisfies the conditions we discussed earlier.

1. **Normalization**: if we consider the entire sample space, i.e., $A = \Omega$, we have:

   $$P(\Omega) = \frac{|\Omega|}{|\Omega|} = 1$$

   This means that the probability of getting either heads or tails in a fair coin toss is 1.
2. **Additivity**: since there are only two possible outcomes (heads or tails) in the sample space, any two events are necessarily disjoint. Therefore, the additivity condition holds trivially.

In summary, the probability function accurately reflects the inherent fairness of the experiment, where each outcome is equally likely.

Example: Biased Die

Consider a biased six-sided die with probabilities $P(\{i\}) = \frac{i}{21}$ for $i = 1, 2, \ldots, 6$. This means that the probability of rolling a 1 is $\frac{1}{21}$, of rolling a 2 is $\frac{2}{21}$, and so on, up to the probability of rolling a 6 which is $\frac{6}{21}$. The probability function P defined on single outcomes (i.e., on the individual faces of the die) extends to all events using the principle of additivity. That is, for any event A (a subset of $\{1, 2, 3, 4, 5, 6\}$), the probability of A is given by

$$P(A) = \sum_{i \in A} P(\{i\}).$$

For example, let $\mathcal{O} = \{1, 3, 5\}$ be the event that the outcome is an odd number. Then:

$$P(\mathcal{O}) = P(\{1\}) + P(\{3\}) + P(\{5\}) = \frac{1}{21} + \frac{3}{21} + \frac{5}{21} = \frac{9}{21} = \frac{3}{7}.$$

This shows how probabilities of more complex events are built up from the probabilities of individual outcomes.

1. **Normalization**: Let's calculate the probability for the entire sample space:

   $$P(\Omega) = P(\{1\})+P(\{2\})+P(\{3\})+P(\{4\})+P(\{5\})+P(\{6\}) = \frac{1}{21}+\frac{2}{21}+\frac{3}{21}+\frac{4}{21}+\frac{5}{21}+\frac{6}{21} = \frac{21}{21} = 1.$$

   This demonstrates that the probabilities sum up to 1, confirming the normalization condition.

2. **Additivity**: This property holds because the function is defined on one-element subsets (single outcomes), and extended by the additivity property to unions of events. For instance, $P(\{1\} \cup \{2\} \cup \{3\}) = P(\{1\}) + P(\{2\}) + P(\{3\}) = \frac{1}{21} + \frac{2}{21} + \frac{3}{21} = \frac{6}{21}$.

While this probability function results in a die that is biased towards higher values, this does not invalidate it as a probability function. It simply means that the die is not fair, and the outcomes are more likely to be on the higher end.

In summary, the function $P(\{i\}) = \dfrac{i}{21}$ for $i = 1, 2, \ldots, 6$ is indeed a valid probability function, even though it represents a biased die.

Remark. The extensive explanation here was for clarity You won't need to provide such detailed answers on upcoming assignments.

As already discussed, sample spaces can be other than finite or discrete sets. In these cases, it is more common to use the term *probability distribution* instead of probability function.

Example: a Continuous Probability Distribution

A foretaste of what's to come: let's explore the fascinating realm of continuous probability distributions. Consider the function $f(x) = 3x^2$, and for any $S \subseteq [0, 1]$, define the probability of S by

$$P(S) = \int_S 3x^2 \, dx$$

.

Let's go through the checks:

1. **Non-Negativity.**

   $$f(x) = 3x^2 \geq 0$$

   for all x in the interval $[0, 1]$. This condition is satisfied.

2. **Normality.**

   $$\int_0^1 3x^2 \, dx = x^3 \Big|_0^1 = 1 - 0 = 1$$

   The integral evaluates to 1, indicating that the function is normalized.

Remark. The German mathematician Gottfried Wilhelm Leibniz introduced the integral symbol $\int$ to represent 'summa' or 'total'. It is like a shortcut for adding up lots of tiny pieces. This helps us to transit from adding values over 'discrete domains' to finding totals over 'continuous domains'.

3. **Additivity.** For any two disjoint subset $S_1, S_2 \subset [0,1]$, we have $\int\limits_{S_1 \cup S_2} f(x)dx = \int\limits_{S_1} f(x)dx + \int\limits_{S_2} f(x)dx$ by definition of integral.

Nonexamples

1. Consider a sample space of three possible outcomes: $\Omega = \{A, B, C\}$. We define a probability function P as follows:

$$\begin{aligned} P(\{A\}) &= 0.3, \\ P(\{B\}) &= 0.4, \\ P(\{C\}) &= 0.5. \end{aligned}$$

   In this case, the probabilities assigned do not add up to 1:

$$P(\Omega) = P(\{A\}) + P(\{B\}) + P(\{C\}) = 0.3 + 0.4 + 0.5 = 1.2.$$

   Since the total probability is greater than 1, this violates the normalization condition.

2. Consider a sample space of two outcomes: $\Omega = \{T, F\}$, We define a function $P : \Omega \to \mathbb{R}$ as follows:

$$\begin{aligned} P(\{T\}) &= 0.4, \\ P(\{F\}) &= 0.5, \\ P(\{T, F\}) &= 1. \end{aligned}$$

   In this case, the probabilities assigned violate the additivity condition. If we take $A = \{T\}$ and $B = \{F\}$, then A and B are disjoint sets (meaning they cannot both happen at the same time), but:

$$P(A \cup B) = P(\{T, F\}) = 1 \neq P(\{T\}) + P(\{F\}) = 0.4 + 0.5 = 0.9.$$

   This shows that the additivity condition is not satisfied for this function.

Question. How many probability functions are there on the sample space $\Omega = \{T, F\}$?

As we discussed, a probability function on a finite sample space is completely determined by its values on singleton events. In this case, $\Omega = \{T, F\}$, so it is enough to define:

$$0 \leq P(\{T\}) \leq 1 \quad \text{and} \quad 0 \leq P(\{F\}) \leq 1.$$

However, the normality axiom of probability requires that:

$$P(\{T, F\}) = P(\{T\}) + P(\{F\}) = 1.$$

This means that once we choose a value for $P(\{T\})$, the value of $P(\{F\})$ is determined automatically by:

$$P(\{F\}) = 1 - P(\{T\}).$$

Therefore, the set of all probability functions on Ω is in one-to-one correspondence with the interval $[0, 1]$. Each number $a \in [0, 1]$ defines a valid probability function via:

$$P(\{T\}) = a, \quad P(\{F\}) = 1 - a.$$

Thus, there are infinitely many such probability functions, one for each real number in the interval $[0, 1]$.

Lecture 3
Exclusive Offer: All-Inclusive Probability Journey

The inclusion-exclusion principle is a fundamental concept in combinatorics and probability theory. It provides a systematic way to count or calculate probabilities of events that involve the union and intersection of sets.

Inclusion-Exclusion Principle for Two Sets

Imagine a scenario in which a company has two open positions. Specifically, 30 individuals have applied for the first position, while 25 have applied for the second. Interestingly, 10 candidates have shown interest in both positions. The director of the company would like to determine the overall number of applicants.

To address this, let's define A as the set of applicants for the first position and B as the set for the second position. At first glance, it might seem logical to simply add the number of applicants for A and B together. However, this approach would lead to an inaccurate count, as those 10 applicants who applied for both positions would be counted twice. Therefore, to arrive at the correct answer, we need to account for this overcounting. This can be achieved by subtracting the number of applicants who applied for both positions from the total count of applicants for both positions. This gives us the accurate total number of unique applicants, which equals 45.

In mathematical terms:

$$|A \cup B| = |A| + |B| - |A \cap B| = 30 + 25 - 10 = 45.$$

Now that we've successfully determined the total number of unique applicants in this scenario, let's take a moment to understand the broader principle behind this.

This process of correcting for overcounting by subtracting the number of shared elements is a fundamental idea in mathematics. It's known as the Inclusion-Exclusion Principle. This principle provides a powerful tool for dealing with overlapping sets.

In the case of two sets, like our applicants for positions A and B, the Inclusion-Exclusion Principle can be expressed as:

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

This formula allows us to find the total number of unique elements in two sets, even when there is overlap.

Example: Rolling a Die

Let's apply this principle to a simple example. Suppose we roll a fair six-sided die.

Let A be the event of rolling an even number, and B be the event of rolling a number greater than 3.

$$\begin{aligned} |A| &= 3 \quad \text{(numbers 2, 4, 6)} \\ |B| &= 3 \quad \text{(numbers 4, 5, 6)} \\ |A \cap B| &= 2 \quad \text{(only 4 and 6 satisfy both)} \end{aligned}$$

Using the inclusion-exclusion principle:

$$|A \cup B| = |A| + |B| - |A \cap B| = 3 + 3 - 2 = 4$$

So, there are 4 outcomes in which either an even number or a number greater than 3 occurs.

Applications to Probability Theory

The inclusion-exclusion principle finds extensive applications in probability theory. It enables us to calculate probabilities of complex events by considering the union and intersection of simpler events.

Example. Let's revisit the example of rolling a fair die. We want to find the probability of rolling a number greater than 3 or rolling an even number.

Directly, the set of outcomes satisfying this condition consists of 2, 4, 5, and 6. So, the probability is:

$$P(\{2,4,5,6\}) = \frac{4}{6} = \frac{2}{3}.$$

Now, let's use the inclusion-exclusion principle. Here, $A = \{2,4,6\}$ is the event of rolling an even number, and $B = \{4,5,6\}$ is the event of rolling a number greater than 3.

$$\begin{aligned} P(A) &= \frac{3}{6} = \frac{1}{2} \\ P(B) &= \frac{3}{6} = \frac{1}{2} \\ P(A \cap B) &= \frac{2}{6} = \frac{1}{3} \end{aligned}$$

Using the inclusion-exclusion principle:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{2} - \frac{1}{3} = \frac{2}{3}.$$

So, we can see that both methods yield the same result. The probability of rolling a number greater than 3 or an even number is $\frac{2}{3}$.

Example. Suppose you have a standard deck of 52 playing cards. What is the probability of drawing either a red card or a face card?

Directly calculating this probability can be a bit tricky. There are 26 red cards in the deck (hearts and diamonds), and there are 12 face cards (4 kings, 4 queens, and 4 jacks). However, we can't simply add these probabilities together because there are 6 cards (2 red jacks, 2 red queens, and 2 red kings) that are both red and face cards.

To find the probability, we will use the inclusion-exclusion principle. Let A be the event of drawing a red card, and B be the event of drawing a face card.

$$\begin{aligned} P(A) &= \frac{26}{52} \text{ (since half of the cards are red);} \\ P(B) &= \frac{12}{52} \text{ (since there are 12 face cards in total);} \\ P(A \cap B) &= \frac{6}{52} \text{ (since there are 6 cards that are both red and face cards).} \end{aligned}$$

Using the inclusion-exclusion principle:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{26 + 12 - 6}{52} = \frac{32}{52} = \frac{8}{13}.$$

So, the probability of drawing either a red card or a face card is approximately 0.615 or 61.5%. This example demonstrates how the inclusion-exclusion principle is essential in situations where events overlap and straightforward calculation isn't obvious.

Real-Life Example: Birthday Parties

Consider a group of friends who have the following common interests in birthday parties:

- A: Friends who enjoy outdoor activities.
- B: Friends who enjoy indoor activities.

Out of the group of 30 friends, we assume that every friend enjoys at least one activity.

- 15 friends enjoy outdoor activities ($|A| = 15$).

- 20 friends enjoy indoor activities ($|B| = 20$).
- 10 friends enjoy both indoor and outdoor activities ($|A \cap B| = 10$).

We want to calculate the probability of inviting a friend who enjoys either outdoor or indoor activities to a birthday party.

Using the inclusion-exclusion principle:

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

$$= \frac{|A|}{|\text{Total friends}|} + \frac{|B|}{|\text{Total friends}|} - \frac{|A \cap B|}{|\text{Total friends}|}$$

$$= \frac{15}{30} + \frac{20}{30} - \frac{10}{30} = \frac{25}{30} = \frac{5}{6} \approx 0.833.$$

So, the probability of inviting a friend who enjoys either outdoor or indoor activities to a birthday party is approximately 0.833 or 83.3%.

Remark. It is important to assume that everyone enjoys at least one activity because it ensures that the total probability of all possible events is equal to 1. In other words, it guarantees that there are no friends left out in this scenario.

Proof Idea for Two Sets

To prove the inclusion-exclusion principle for two sets, we start by considering the sizes of A, B, and their intersection $A \cap B$.

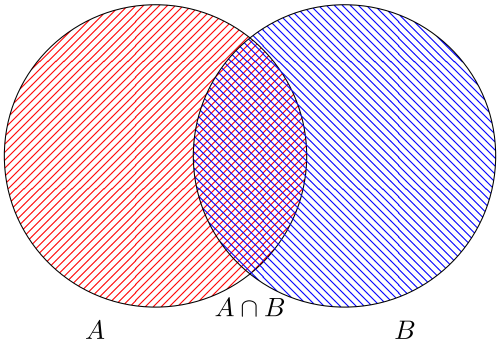
$$\begin{aligned} |A| &= \text{number of elements in } A \\ |B| &= \text{number of elements in } B \\ |A \cap B| &= \text{number of elements in both } A \text{ and } B \end{aligned}$$

When we sum the sizes of A and B, we count the elements in $A \cap B$ twice, so we need to subtract $|A \cap B|$ to correct for this double-counting.

$$|A| + |B| - |A \cap B|$$

This accounts for all elements in either A or B without double-counting.

In the Venn diagram below, we illustrate sets A and B, along with their intersection $A \cap B$. Notice that $A \cap B$ is shaded with a line segment pattern in both A and B to emphasize that it is counted twice:



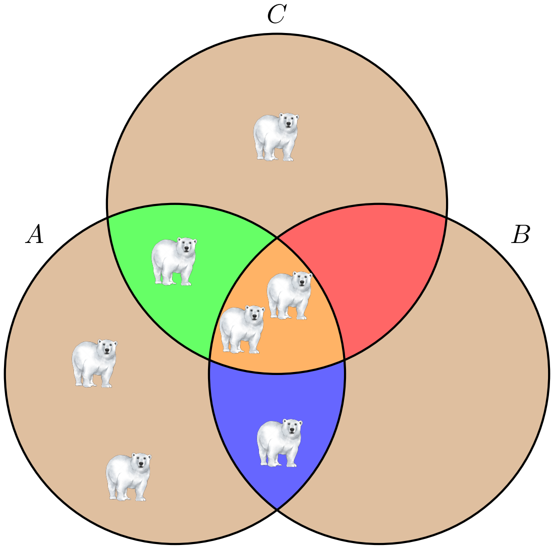
Generalization to Three or More Sets

Note: This section contains additional material for those interested in further exploring the topic.

The inclusion-exclusion principle can be extended to more than two sets. For three sets A, B, and C, it is given by the formula

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$

Example. Imagine we are studying the habitats of polar bears. In this context, A, B and C represent distinct environments where these bears reside. For instance, A may be a region with 6 polar bears, B another area with 3 bears, and C a third area with 4 bears. Now, it's not uncommon for a polar bear to traverse different habitats. We could humorously attribute this to the bears' keen instinct for locating the richest salmon feeding grounds. When faced with a choice between two thriving areas, our bear may gravitate towards the one with the most abundant salmon population. Suppose we are interested in computing the total number of bears, taking into account that there are 3 bears migrating between habitats A and B, 3 bears between A and C, 2 bears alternating between B and C, and finally, 2 intrepid bears constantly moving around all three habitats:



Using the inclusion-exclusion principle for three sets we get

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C| = 6 + 3 + 4 - 3 - 3 - 2 + 2 = 7.$$

This means that there are 7 polar bears in total across all three habitats. This example demonstrates how the inclusion-exclusion principle helps us count elements in the union of multiple sets, taking into account their intersections.

When applying the inclusion-exclusion principle, think of it as an iterative process. At each stage, we add or subtract the size of intersections of increasing order to ensure that each element is counted exactly once. This systematic approach is especially useful when dealing with overlapping sets in complex counting problems.

In general, for n sets $A_1, A_2, \ldots, A_n$, the inclusion-exclusion principle is given by:

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{i=1}^{n} |A_i| - \sum_{i<j} |A_i \cap A_j| \quad + \sum_{i<j<k} |A_i \cap A_j \cap A_k| - \ldots + (-1)^{n-1} |A_1 \cap A_2 \cap \ldots \cap A_n|$$

This formula expresses the total number of elements in the union of n sets in terms of the cardinalities (number of elements) of all possible intersections among those sets.

Remark. The name "inclusion-exclusion" reflects the alternating nature of the formula: we first include the sizes of individual sets, then exclude overcounted pairwise intersections, then re-include triplewise intersections, and so on.

Lecture 4
Multiplication Rule and Counting Principles

Our focus today centers around the multiplication rule and counting principles, crucial elements in the world of probability. These tools serve as the cornerstone of probability theory with finite sample spaces, providing us with the means to address a diverse array of probabilistic situations.

The Multiplication Rule

When we have multiple experiments, each with its own set of outcomes, we can calculate the total number of possible outcomes by multiplying the number of outcomes for each experiment.

🥪 Building a Sandwich 🥪

Imagine you have 4 types of bread, 5 types of fillings, and 3 types of condiments to choose from. You would like to create a sandwich with 1 piece of bread, one filling, and one condiment. To find the total number of possible sandwich combinations, you need to consider all the choices you have for each component. You have:

- 4 options for the type of bread,
- 5 options for the filling, and
- 3 options for the condiment.

To find the total number of combinations, you multiply these options together:

$$4 \cdot 5 \cdot 3 = 60.$$

This means you have a total of 60 unique sandwich combinations to choose from.

Example: Drawing Cards Without Replacement

Imagine we have a standard deck of 52 cards. If we draw three cards without replacement, the number of possible outcomes can be calculated as:

$$52 \cdot 51 \cdot 50$$

Each draw affects the available choices for the next draw, giving us a total of $132,600$ possible combinations.

General Formula

Let's say we have k experiments, where the first experiment has m_1 outcomes, the second experiment has m_2 outcomes, and so on, up to the k-th experiment with m_k outcomes. The total number of possible outcomes is given by:

$$m_1 \cdot m_2 \cdot \ldots \cdot m_k$$

This formula is known as the Multiplication Rule.

Number of Arrangements & Factorial

Suppose you have 5 different books and you want to arrange them on a shelf. The first spot can be occupied by any of the 5 books. So, there are 5 choices for the first spot.

Now, for the second spot, you have 4 remaining books to choose from, since you've already placed one book in the first spot.

Similarly, for the third spot, there are 3 remaining choices, and so on.

To find the total number of arrangements, you multiply the number of choices for each spot:

$$5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120.$$

This can be expressed as 5!, which is the factorial of 5.

Definition. The **factorial** of a non-negative integer n, denoted as $n!$, is the product of all positive integers up to n, with the convention that $0! = 1$. Mathematically, it is defined as:

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1.$$

Example. Consider the word "WORK". If we want to rearrange its letters, there are 4! ways to do so:

$$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$$

This means there are 24 different permutations of the letters in the word "WORK".

Permutations: Order Matters

The number of ways to arrange k items from a set of n distinct items, considering the order, is denoted as $P(n, k)$ and can be calculated using the formula:

$$P(n, k) = \frac{n!}{(n-k)!}$$

This formula represents permutations, where the order of arrangement matters.

Explanation

To understand why this formula computes the number of arrangements, let's break it down:

- $n!$ represents the number of ways to arrange all n items without any restrictions. This is because there are n choices for the first item, $n-1$ choices for the second item, and so on, down to 1 choice for the last item.

- $(n-k)!$ represents the number of ways to arrange the remaining $(n-k)$ items after k items have been selected and arranged. This accounts for the fact that we are considering k items and their order, so we don't want to count the arrangements of the unselected items.

Dividing $n!$ by $(n-k)!$ eliminates the arrangements of the unselected items, leaving us with the number of arrangements of the k selected items.

Therefore, $P(n, k)$ gives us the total number of permutations of k items chosen from n distinct items, considering the order of arrangement.

Example. In a basketball tournament, there are 10 players competing for 5 spots on the starting lineup, and each spot has a specific position. How many different starting lineups can be formed?

Since each position in the lineup is distinct, the order in which the players are assigned matters. Thus, we are selecting and arranging 5 players out of 10, which is a permutation.

The number of different starting lineups is:

$$P(10, 5) = \frac{10!}{(10-5)!} = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 = 30240.$$

So, there are 30240 different possible starting lineups.

Combinations: Order Doesn't Matter

The number of ways to choose k items from a set of n distinct items, without considering the order, is denoted as $C(n,k)$ or $\binom{n}{k}$, and can be calculated using the formula:

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

This formula represents combinations, where the order of selection doesn't matter.

Explanation

To understand why this formula counts the right number, let's refer back to the concept of permutations discussed earlier.

In permutations, we calculated $P(n,k)$ to find the number of arrangements of k items from n distinct items where order matters. This means that if we had the same k items, there would be $k!$ different ways to arrange them.

However, in combinations, we want to find the number of ways to choose k items without considering the order. This means that for each combination of k items, there are $k!$ different arrangements that are equivalent. Therefore, to get the correct count of combinations, we divide $P(n,k)$ by $k!$ to eliminate the $k!$ duplicate arrangements.

This leads us to the formula for combinations, $C(n,k) = \dfrac{n!}{k! \cdot (n-k)!}$, which accurately counts the number of ways to choose k items from n distinct items, without considering the order.

Example. In a Venice Beach basketball tournament, there are 10 players competing for 5 spots on the starting lineup. However, the positions on the lineup are not fixed or assigned (e.g., it's not predetermined who plays point guard, center, etc.). How many different starting lineups can be formed?

This time any group of 5 players is considered the same. Since the order of selection does not matter, this is a combination:

$$\binom{10}{5} = \frac{10!}{5! \cdot 5!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 252.$$

So, there are 252 different possible starting lineups.

Example. Suppose you have 8 people and want to form a committee of 3 members.

(a) If there are no specific roles assigned (e.g., all members are equal), then the order of selection does not matter.

This is a combination:

$$\binom{8}{3} = \frac{8!}{3! \cdot 5!} = \frac{8 \cdot 7 \cdot 6}{3 \cdot 2 \cdot 1} = 56.$$

There are 56 different ways to choose a committee of 3 people from a group of 8.

(b) Now suppose you want to select a President, Vice President, and Treasurer from a group of 8 people. Since each role is different, the order in which people are selected matters.

This is a permutation:

$$P(8,3) = \frac{8!}{(8-3)!} = 8 \cdot 7 \cdot 6 = 336.$$

There are 336 different ways to assign 3 distinct roles to 3 people from a group of 8.

Counting Words Formed from Given Letters

To count how many distinct "words" (strings of letters) can be formed from a given set of letters, we need to consider whether the letters are all distinct or whether some letters are repeated.

Example. 1. How many different words can be formed using all the letters in the word '`abcd`'?

Since all 4 letters are distinct, the number of permutations is:

$$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24.$$

So, there are 24 different arrangements of the letters in '`abcd`'.

2. Consider the word '`acdc`'. Notice that the letter `c` appears twice.

To count the number of distinct permutations, we divide by the factorial of repeated letters:

$$\frac{4!}{2!} = \frac{24}{2} = 12.$$

So, there are 12 different arrangements of the letters in '`acdc`'.

Remark. Yes, `ACDC` is also the name of a famous Australian rock band! The name stands for *Alternating Current / Direct Current*, a reference to electric power systems.

3. How many different words can be formed from the letters in '`abracadabra`'?

The word '`abracadabra`' has 11 letters in total. The frequency of each letter is:

- `a` appears 5 times,
- `b` appears 2 times,
- `r` appears 2 times,
- `c` and `d` appear 1 time each.

So the number of distinct permutations is:

$$\frac{11!}{5! \cdot 2! \cdot 2! \cdot 1! \cdot 1!} = 83160.$$

Thus, there are 83160 distinct words that can be formed from '`abracadabra`'.

Remark. The formula used above is a generalization of the binomial coefficient, called a **multinomial coefficient**. It counts the number of distinct permutations of objects when some of them are indistinguishable. The binomial coefficient $\binom{n}{k}$ is just a special case of this when there are only two types of items.

Permutations with Replacement

When the order of selection matters and items can be chosen more than once, we use the formula for permutations with replacement. If we have n choices and we make k selections with replacement, the number of permutations is n^k.

Example. Consider a 4-digit PIN where each digit can be any number from 0 to 9. The number of possible PINs is:

$$10^4 = 10,000$$

So, there are 10000 different possible 4-digit PINs.

Lecture 5

General and Conditional Probability

As we briefly touched upon earlier, understanding counting principles is paramount when it comes to tackling probability on finite sample spaces. These principles provide the fundamental tools needed to navigate through various probability scenarios. Let's apply these concepts to a range of examples.

Example. An urn contains 4 red balls, 3 green balls, and 2 blue balls. If we draw 3 balls at random **without replacement**, what is the probability of getting **exactly 2 red balls**?

There are $4 + 3 + 2 = 9$ total balls in the urn. The number of ways to choose 3 balls from 9 is:

$$\binom{9}{3} = \frac{9!}{3! \cdot 6!} = 84.$$

To get exactly 2 red balls:

$$\binom{4}{2} = 6 \quad \text{(ways to choose 2 red balls out of 4)}.$$

The remaining 1 ball must be non-red — either green or blue. There are $3 + 2 = 5$ such balls:

$$\binom{5}{1} = 5 \quad \text{(ways to choose 1 non-red ball out of 5)}.$$

So, the total number of favorable outcomes is:

$$\binom{4}{2} \cdot \binom{5}{1} = 6 \cdot 5 = 30.$$

Finally, we compute the probability under consideration:

$$P(\text{exactly 2 red balls}) = \frac{30}{84} = \frac{5}{14}.$$

Example. If you draw 3 cards from a standard deck of 52 cards, what is the probability that all 3 are hearts?

There are $\binom{13}{3}$ ways to choose 3 hearts out of the 13 available. The total number of ways to choose any 3 cards from the deck is $\binom{52}{3}$. Therefore, the probability is:

$$P(\text{3 hearts}) = \frac{\binom{13}{3}}{\binom{52}{3}} = \frac{13 \cdot 12 \cdot 11}{52 \cdot 51 \cdot 50}.$$

Remark. Alternatively, suppose we draw the cards one by one, without replacement. The probability that the first card is a heart is $\frac{13}{52}$. Given that, the probability that the second card is also a heart is $\frac{12}{51}$, and then $\frac{11}{50}$ for the third. So the probability is:

$$P(3 \text{ hearts}) = \frac{13}{52} \cdot \frac{12}{51} \cdot \frac{11}{50}.$$

Both methods yield the same result, though they differ in approach. The combination method counts favorable and total outcomes directly, while the sequential method follows the probability step by step. The latter is often helpful for building intuition, especially when reasoning about conditional probabilities.

Example. Suppose you roll two fair six-sided dice. What is the probability that the sum of the numbers showing on the two dice is exactly 7?

To solve this, we analyze all possible outcomes. Each die has 6 sides, so there are $6 \cdot 6 = 36$ possible outcomes when rolling two dice.

We can visualize the outcomes in the table below, where a is the number rolled on the first die and b is the number on the second die. The entries represent the sum $a + b$.

a / b	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

The red entries show the outcomes where the sum is 7. These are the pairs:

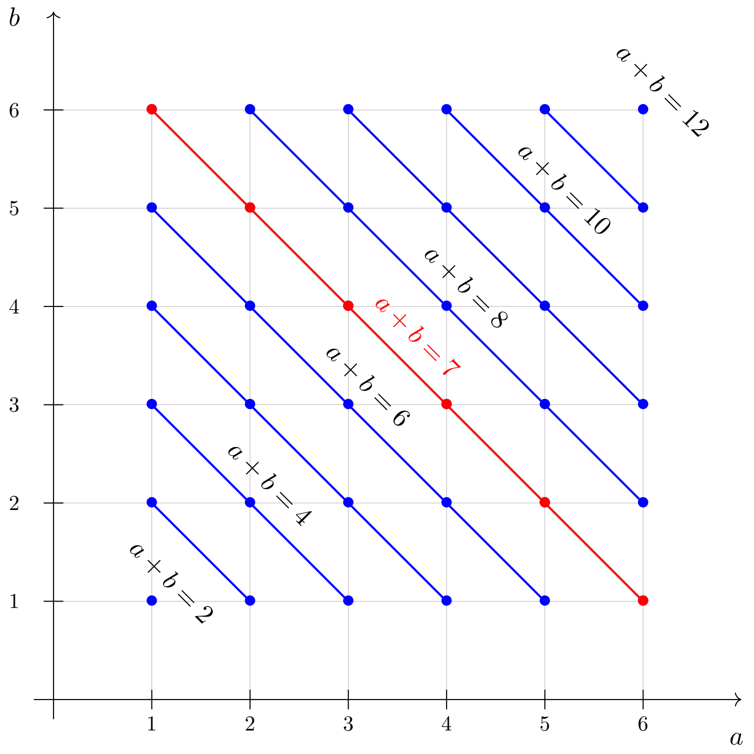
$$(1, 6),\ (2, 5),\ (3, 4),\ (4, 3),\ (5, 2),\ (6, 1).$$

So there are 6 favorable outcomes.

Since all 36 outcomes are equally likely, the probability of getting a sum of 7 is:

$$P(a + b = 7) = \frac{6}{36} = \frac{1}{6}.$$

Remark. It is useful to think of the condition $a + b = k$ as the equation of a line on the ab-plane. The number of favorable outcomes corresponds to the number of integer lattice points lying on the line segment $a + b = k$ within the square $1 \leq a, b \leq 6$.



Notice the symmetry about the sum of 7. For any integer $1 \leq k \leq 5$, we have:

$$P(a+b=7-k) = P(a+b=7+k).$$

For instance, $P(a+b=5) = P(a+b=9) = \frac{4}{36}$, $\quad P(a+b=4) = P(a+b=10) = \frac{3}{36}$, $\quad$ and so on.

Remark. An optional but instructive exercise is to generalize this setup to the case where you roll three fair six-sided dice. Let a, b, c be the numbers shown on the three dice. We are interested in computing the probability

$$P(a+b+c=k)$$

for various values of k.

There are $6^3 = 216$ possible outcomes when rolling three dice. Each outcome corresponds to a triple (a, b, c) where $1 \leq a, b, c \leq 6$. Geometrically, these outcomes form the set of integer lattice points in the 3D cube $[1,6] \times [1,6] \times [1,6]$.

For a fixed value of k, the condition $a+b+c=k$ defines a plane in 3D space. Thus, computing $P(a+b+c=k)$ amounts to counting the number of lattice points in the cube that lie on the plane $a+b+c=k$, and dividing that count by 216:

$$P(a+b+c=k) = \frac{\#\{(a,b,c) \in [1,6] \times [1,6] \times [1,6] \mid a+b+c=k\}}{216}$$

Question. For which value(s) of k is this probability the largest?

Hint. The minimum possible sum is $1+1+1=3$, and the maximum is $6+6+6=18$. Is there a symmetry around the midpoint $(3+18)/2 = 10.5$? If so, the most probable sums should be those closest to 10.5. Geometrically, this reflects a central symmetry with respect to the plane $a+b+c=10.5$, which splits the cube of outcomes evenly. This is analogous to the symmetry we observed earlier for two dice, where the number of outcomes with fixed sum $a+b=k$ was symmetric around the line $a+b=(2+12)/2=7$. In both cases, the most probable outcomes correspond to sums nearest the midpoint of the possible range.

Using Complements to Simplify Probability Calculations

One very useful strategy in probability is working with complements. If $A \subseteq \Omega$ is an event, then:

$$P(A) = 1 - P(A^c),$$

where A^c is the complement of A.

Remark. The above identity follows from the facts that:

$$A \cap A^c = \varnothing \quad \text{and} \quad A \cup A^c = \Omega,$$

implying that for any probability function P,

$$P(\Omega) = P(A \cup A^c) = P(A) + P(A^c) = 1.$$

This identity is especially useful when it is difficult to compute $P(A)$ directly, but easier to find $P(A^c)$.

Example. A basket contains 50 fruits: 30 apples and 20 oranges. Suppose 15 fruits are selected at random without replacement. What is the probability that **at least 2 oranges** are chosen?

Instead of directly computing the probability of getting 2 or more oranges, we apply the *complement strategy*:

$$P(\text{at least 2 oranges}) = 1 - P(\text{fewer than 2 oranges}) = 1 - (P(\text{0 oranges}) + P(\text{1 oranges})).$$

We compute each of these using probabilities using the multiplication rule:

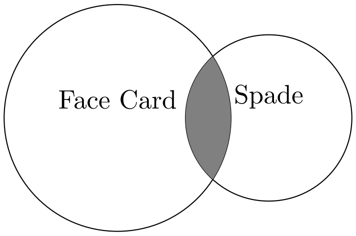
$$\begin{aligned} P(\text{0 oranges}) &= \frac{\binom{30}{15}\binom{20}{0}}{\binom{50}{15}} = \frac{\binom{30}{15}}{\binom{50}{15}}, \\ P(\text{1 orange}) &= \frac{\binom{30}{14} \cdot 20}{\binom{50}{15}}. \end{aligned}$$

Therefore, $P(\text{at least 2 oranges}) = 1 - \frac{\binom{30}{15}}{\binom{50}{15}} - \frac{\binom{30}{14}\cdot 20}{\binom{50}{15}}$.

This approach avoids computing the many more cases involved in directly summing probabilities for 2 through 15 oranges.

Conditional Probability

Consider a standard deck of 52 playing cards. You draw one card at random, but before revealing it, you are informed that the card is a face card (jack, queen, or king). With this knowledge in hand, what is the probability that the card is a spade?



Face Card $\cap$ Spade

Intuitively, knowing that you have a face card should increase the likelihood of having a spade. This is where conditional probability comes into play.

Definition. The **conditional probability** of event A given that event B has already occurred, denoted as $P(A|B)$, is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ where } P(B) \neq 0.$$

Example. The 'Four of a Kind' is a rare and formidable hand in poker, featuring four cards of the same rank.

1. **Four of a Kind (Four Aces).**

   To find the probability of getting four Aces in a five-card hand, we first calculate the number of ways to choose four Aces from the deck (which is $\binom{4}{4} = 1$) and the number of ways to choose the fifth card (which is $\binom{48}{1} = 48$). The total number of five-card hands is $\binom{52}{5} = 2,598,960$. Therefore, the probability is

   $$P(\text{Four Aces}) = \frac{\binom{4}{4} \cdot \binom{48}{1}}{\binom{52}{5}} \approx 0.000018 \approx 0.0018\%.$$

2. **Conditional Probability (Given One Ace).**

   If we already know that one of the cards is an Ace, there are $\binom{3}{3} \cdot \binom{48}{1}$ ways to choose the remaining three cards from the three remaining Aces and the other 48 cards. The total number of four-card hands with one Ace is $\binom{51}{4} = 249,900$. Therefore, the conditional probability is

   $$P(\text{Four Aces} \,|\, \text{One Ace}) = \frac{\binom{3}{3} \cdot \binom{48}{1}}{\binom{51}{4}} \approx 0.00019 \approx 0.019\%.$$

3. **Conditional Probability (Given Two Aces).**

   If we know that both cards in hand are Aces, the conditional probability is

   $$P(\text{Four Aces} \,|\, \text{Two Aces}) = \frac{\binom{2}{2} \cdot \binom{48}{1}}{\binom{50}{3}} \approx 0.0024 \approx 0.24\%.$$

Now, let's address a crucial question that often arises when dealing with conditional probabilities.

Question. Let A and B be two events. Why the probabilities $P(A|B)$ (probability of A given B) and $P(B|A)$ (probability of B given A) differ even though they both involve conditional probabilities??

A very important observation. The formulas for computing conditional probabilities are similar, but **NOT** the same:

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

while

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

The numerators in both formulas coincide and are equal to the probability of occurrence of both events A and B. The difference between the two formulas is that the former formula 'measures the proportion of the intersection relative to B', while the latter 'measures the proportion of the intersection relative to A'.

Here is an example, where the difference between the two probabilities is strikingly evident.

Example. Consider two events:

- A: a person is a natural born citizen of the United States;
- B: a person is the president of the United States.

Remark. Recall that 'No Person except a natural born Citizen, or a Citizen of the United States, at the time of the Adoption of the Constitution, shall be eligible to the Office of President', so $B \subset A$ is a subset.

Then $P(B|A)$ stands for the probability that a person who is a natural born citizen of the United States, is the president of the country, while $P(A|B)$ is the probability that someone who is a president of the United States is a natural born citizen of the United States. Notice that $P(B|A) < \frac{1}{10^7}$ is incredibly small, while $P(A|B) = 1$ (see the remark above).

Notice that the formulas above give rise to two expressions for the probability of intersection of two events:

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

Let's take a look at some more examples.

Example. 1. There is an amazing basketball team ABT and the best player on that team is the player JM.[1] Consider two events:

- A: team ABT wins the game;
- B: player JM participates in the game.

Then $P(B|A)$ stands for the probability that JM takes part in the game that ABT team wins, while $P(A|B)$ is the probability that ABT team wins the game, when JM plays.

2. Consider two events:

- A: a child loves snickers bars ;
- B: a child's parent buys the child a snickers bar .

Now $P(A|B)$ stands for the probability that a child loves snickers bars given his (her) parent bought him (her) one, while $P(B|A)$ is the probability that a child's parent bought the child a snickers given that he (she) loves those.

3. Consider two events:

- A: a person has disease $\mathfrak{h}$;
- B: the result of diagnostic test for disease $\mathfrak{h}$ is correct.

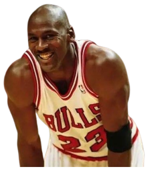
This time $P(B|A)$ stands for the probability that a person, who has the disease $\mathfrak{h}$, tested positive (the test detected the disease correctly), while $P(A|B)$ stands for the probability that the outcome of the test was correct, and the person, who took the test, has the disease.

Interpretation. In this context, it's crucial to understand the implications of false positives and false negatives. A false positive occurs when the test indicates a disease is present, but it is not. This can lead to unnecessary worry and further testing. On the other hand, a false negative occurs when the test indicates no disease, but it is actually present. This can delay necessary treatment. Therefore, interpreting test results should be done cautiously, considering both types of errors.

Law of Total Probability

Let $B_1, B_2, \ldots, B_n$ be a **partition** of the sample space Ω:

- the sets $B_1, \ldots, B_n$ are pairwise disjoint: $B_i \cap B_j = \varnothing$ for $i \neq j$;

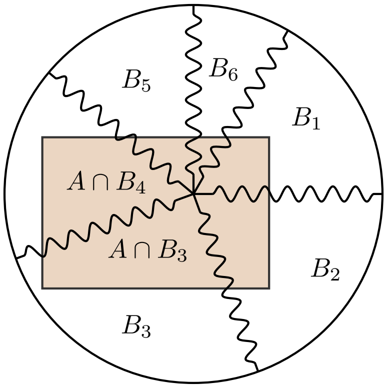
[1]  : Feel free to read from right to left.

- their union covers the entire sample space: $\bigcup_{i=1}^{n} B_i = \Omega$.

Then for any event $A \subseteq \Omega$, the **Law of Total Probability** states:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i).$$

This formula allows us to compute the probability of an event A by breaking it down over a complete set of mutually exclusive cases $B_1, \ldots, B_n$, even when we do not know which case actually occurs.



In the diagram above:

- the circle represents the full sample space Ω;
- it is partitioned into six disjoint regions, each corresponding to one of the events $B_1, \ldots, B_6$;
- the shaded rectangle represents the event A, which overlaps with several of the B_i's;
- each intersection $A \cap B_i$ represents the part of A that lies within B_i.

Since the events $A \cap B_1, A \cap B_2, \ldots, A \cap B_6$ are disjoint and their union equals A, we can write:

$$P(A) = \sum_{i=1}^{6} P(A \cap B_i) = \sum_{i=1}^{6} P(A \mid B_i) \cdot P(B_i).$$

Lecture 6

Bayes' Formula and Independence

Today, we unravel the intricacies of Bayes' Formula. Like secret keys, this concept unlock doors to a world of informed decision-making. In finance, biology, and even in your daily life, the influence of this formula is profound.

Let A and B be two events in a sample space Ω. Recall that the conditional probability of A given B, denoted $P(A|B)$, is defined as:

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

From this definition, we can express the probability of the intersection of A and B in two equivalent ways:

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

These two expressions are especially useful depending on which conditional probability is more readily available.

In addition, the Law of Total Probability allows us to compute the probability of an event B by conditioning on a partition of the sample space into two disjoint subsets A, and A^c:

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

Substituting this into the expression for $P(A|B)$, we get:

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

This result is known as **Bayes' Theorem** (or Bayes' Formula). This formula allows us to reverse conditional probabilities, calculating $P(A|B)$ from $P(B|A)$ and the prior probabilities of A and A^c.

Example. 1. The amazing basketball team ABT wins 70% of their games. The leading player, JM, participates in 90% of the games the team wins and 25% of the games the team loses (there are no draws).

Suppose you know that JM will participate in the next game. What is the probability that team ABT wins?

Solution. As before, we introduce two events:

- A: team ABT wins the game;
- B: player JM participates in the game.

We are given that $P(A) = 0.7$ (so $P(A^c) = 1 - 0.7 = 0.3$), $P(B|A) = 0.9$ and $P(B|A^c) = 0.25$. The quantity $P(A|B)$ needs to be found.

An application of Bayes' formula gives

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{0.9 \cdot 0.7}{0.9 \cdot 0.7 + 0.25 \cdot 0.3} \approx 89.36\%.$$

2. A recent survey showed that 60% of the time, parents buy their children Snickers bars when they go to grocery stores. Moreover, in 92% of such cases, the child would like to have it. Also, when a parent does *not* buy a Snickers bar, there is still a 70% chance the child would have liked to have it.

What is the probability that a parent bought a Snickers bar, given that the child would like one?

Solution. We introduce the events

- A: a child loves snickers bars ;
- B: a child's parent buys the child a snickers bar .

We are given that $P(B) = 0.6$ (so $P(B^c) = 1 - 0.6 = 0.4$), $P(A|B) = 0.92$ and $P(A|B^c) = 0.7$. The probability $P(B|A)$ needs to be found.

An application of Bayes' formula gives

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} = \frac{0.92 \cdot 0.6}{0.92 \cdot 0.6 + 0.7 \cdot 0.4} \approx 66.35\%.$$

3. A recent study reports that 1% of the student population is infected with Combinatoric Overload disease, a rare mathematical condition denoted by $\mathfrak{c}_\omega$.

Typical symptoms include: nightmares involving factorials, confusing permutations with combinations, and whispering "order matters" during casual conversations.

There exists a diagnostic test for $\mathfrak{c}_\omega$, which detects the infection in 95% of infected individuals and incorrectly flags 5% of healthy individuals as positive.

Joe's latest test result came back positive. What is the probability that he is actually infected?

Solution. We introduce the events

- A: a student has the disease $\mathfrak{c}_\omega$;
- B: the result of diagnostic test for disease $\mathfrak{c}_\omega$ is positive.

We are given that $P(A) = 0.01$ (so $P(A^c) = 1 - 0.01 = 0.99$), $P(B|A) = 0.95$ (the person is infected and the test will detect it, i.e. be positive, in 95% of the cases) and $P(B|A^c) = 0.05$ (the results are incorrect with probability $1 - 0.95 = 0.05$, and the event ’B given A^c’ stands for positive results for non infected individuals[2]). The probability $P(A|B)$ needs to be found.

An application of Bayes’ formula gives

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.05 \cdot 0.99} \approx 16.01\%.$$

Therefore, even with a positive test result, the probability that Joe is actually infected with $\mathfrak{c}_\omega$ is approximately 16.01%.

Remark. This is a classic illustration of how rare-event probabilities can be misunderstood when false positives are not taken into account. Always consult your probability instructor before diagnosing yourself with $\mathfrak{c}_\omega$.

Optional: The Monty Hall Problem

Suppose you are on a game show, and you are given the choice of three doors: behind one door is a car; behind the others, goats. The problem first appeared on the show ’*Let’s Make a Deal*’ hosted by Monty Hall, and it was also featured in the movie ’*21*’.

You pick a door, say number 1, and the host, who knows what’s behind each of the doors, opens another door, say number 3, which has a goat. He then says to you, ’Do you want to pick door number 2?’ Is it to your advantage to switch your choice?

The answer to this question can be given using Bayes’ formula, but I think we have had enough of it for the moment.

Instead, I will present a simple logical argument that shows why altering the initial choice is in your favor (increases the odds of getting the car).

Notice that when you picked the door the first time (at random), with probability 2/3 there was a goat behind it and with probability 1/3 there was a car. Once the host opens the door with a goat, if there was a goat behind the door that you chose initially (a 66.(6)% chance), switching your choice results in finding the car. However, in case your original pick (33.(3)% chance) was the right one, switching results in losing. We conclude that modifying the initial choice gives a 66.(6)% probability of winning, while not altering it gives 33.(3)%.

Question. Suppose you are given the choice of four doors: behind one door is a car; behind the others, goats. Same rules as before. Is it to your advantage to switch your initial choice this time?

Independence of Events

In probability theory, we are often interested in situations where the occurrence of one event does not influence the likelihood of another. In such cases, we say the events are *independent*.

Intuitively, independence means that knowing whether one event occurred gives no additional information about the other. Mathematically, this is first expressed in terms of conditional probability:

$$P(A \mid B) = P(A)$$

That is, the probability of event A occurring, given that B occurred, is just the same as the original probability of A. This leads us to the formal definition.

[2] These are called ’false positives’

Definition. Two events A and B are said to be **independent** if the occurrence of one does not affect the probability of the other. This is equivalent to the following condition:

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

This formula provides a practical way to check whether two events are independent: simply compare the joint probability $P(A \cap B)$ with the product $P(A) \cdot P(B)$.

Example. 1. Consider flipping a fair coin twice. Let A be the event that the first flip results in heads, and let B be the event that the second flip results in tails.

   Since the coin has no memory, the flips are independent. We compute:

$$P(A) = P(B) = \frac{1}{2}, \quad P(A \cap B) = \frac{1}{4}.$$

   As $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, the independence condition holds.

2. Suppose you roll a fair six-sided die twice. Let E be the event that the first roll is an even number (i.e., $2, 4$, or 6), and let B be the event that the second roll is a 6.

   We notice that $E \cap B = \{(2,6), (4,6), (6,6)\}$ and compute:

$$P(A) = \frac{3}{6} = \frac{1}{2}, \quad P(B) = \frac{1}{6}, \quad P(A \cap B) = \frac{3}{36} = \frac{1}{12}.$$

   As $\frac{1}{12} = \frac{1}{2} \cdot \frac{1}{6}$, these two events are independent.

3. Suppose you draw a card from a standard 52-card deck without replacement. Let A be the event that the card is red, and let B be the event that the card is a heart.

   Here, $B \subseteq A$, so the occurrence of B (drawing a heart) guarantees the occurrence of A (drawing a red card), and hence:

$$P(A|B) = 1 \neq P(A) = \frac{26}{52} = \frac{1}{2}.$$

   Therefore, events A and B are **not** independent.

Independence of Many Events

When the occurrence of multiple events is not influenced by the occurrence of any other event, we have independence of many events.

Definition. Events $A_1, A_2, \ldots, A_n$ are said to be **mutually independent** if for every subset $\{i_1, i_2, \ldots, i_k\} \subseteq \{1, 2, \ldots, n\}$, we have

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdots P(A_{i_k}).$$

In other words, the probability of the intersection of any subcollection of the events equals the product of their individual probabilities.

It is important to note that pairwise independence doesn't necessarily imply mutual independence.

Example. Consider three events in a game of throwing two fair dice.

- A: Getting a sum of 7.
- B: The first die rolls a 4.
- C: The second die rolls a 5.

We want to determine if these events are independent.

We know that $P(A) = \frac{1}{6}$, as there are 6 possible ways to get a sum of 7 out of 36 equally likely outcomes. The probability of event B is also $\frac{1}{6}$ since there is one way to get a 4 on the first die. Similarly, the probability of event C is $\frac{1}{6}$.

Now, let's compute the probabilities of the pairwise intersections:

$$\begin{aligned}
P(A \cap B) &= P(\text{Sum is 7 and first die is 4}) = P(\{(4,3)\}) = \frac{1}{36},\\
P(A \cap C) &= P(\text{Sum is 7 and second die is 5}) = P(\{(2,5)\}) = \frac{1}{36},\\
P(B \cap C) &= P(\text{First die is 4 and second die is 5}) = P(\{(4,5)\}) = \frac{1}{36}.
\end{aligned}$$

Since $P(A \cap B) = P(A) \cdot P(B)$, $P(A \cap C) = P(A) \cdot P(C)$, and $P(B \cap C) = P(B) \cdot P(C)$, we can conclude that these events are pairwise independent.

$$P(A \cap B \cap C) = P(\text{Sum is 7, first die is 4, and second die is 5}) = 0.$$

As $P(A \cap B \cap C) = 0 \neq \dfrac{1}{216} = P(A) \cdot P(B) \cdot P(C)$, we see that these events are not mutually independent.

Lecture 7
Random Variables and Probability Distributions

In probability theory, we often encounter situations where we are more interested in assigning numerical values to outcomes, rather than the outcomes themselves. While outcomes like "heads" or "tails" are useful for describing an experiment qualitatively, they are not suitable for performing quantitative analysis. For example, we might want to compute the *average outcome*, or measure the *typical discrepancy from the average*. These considerations require us to work with numbers, not abstract labels. To bridge this gap, we define functions that assign real numbers to the outcomes. These functions are called *random variables*, and they enable us to gain valuable information on probabilistic systems using the tools of algebra and calculus.

Motivating Examples

- **Coin Flip Game.**

  Sample space: $\Omega = \{H, T\}$ (Heads or Tails).

  Rule: If the coin lands on **Heads**, you win \$2. If it lands on **Tails**, you lose \$1. We define a random variable X by
  $$X(H) = 2, \quad X(T) = -1.$$

- **Rolling Two Dice.**

  Sample space: $\Omega = \{(i,j) \mid 1 \le i, j \le 6\}$ consists of 36 ordered pairs.

  Define a random variable $\mathcal{S}$ to be the *sum of the numbers* on the two dice:
  $$\mathcal{S}(i,j) = i + j.$$

  Then the range of $\mathcal{S}$ is $\{2, 3, \ldots, 12\}$.

These examples show how random variables can be used to assign numerical values to outcomes in games, simulations, or real-world processes.

Definition. A **random variable** is a function $X : \Omega \to \mathbb{R}$ that assigns a real number to each outcome in the sample space Ω. That is, a random variable translates outcomes of a probabilistic experiment into numerical values suitable for analysis.[3]

In other words, a random variable is a rule that converts qualitative outcomes into quantitative values.

Types of Random Variables

In the first lecture, we introduced the distinction between *discrete* and *continuous* probability. This terminology applies directly to the range of a random variable.

Definition. A random variable is called **discrete** if its range is finite or *countably infinite.* That means the values it can take can be listed in a sequence (like $1, 2, 3, \ldots$).

A random variable is called **continuous** if its range includes an entire interval or uncountably infinite subset of $\mathbb{R}$.

Example.

1. **Card Game.** Let Ω be the set of all possible draws of 4 cards from a standard shuffled 52-card deck. Define the random variable Y as the number of aces among the 4 cards drawn. Then the range of Y is
$$\text{Range}(Y) = \{0, 1, 2, 3, 4\},$$
which is finite and thus Y is a discrete random variable.

2. **Manufacturing.** Suppose a light bulb factory tests the brightness of its products. Let the sample space Ω be the collection of all bulbs produced, and define the random variable Z to be the brightness (measured in lumens) of a randomly selected bulb. Then Z takes values in an interval such as $[400, 1500] \subset \mathbb{R}$, so it is a continuous random variable.

3. **Call Center.** Let Ω represent all possible one-hour intervals during a typical workday. Define the random variable C to be the number of customer service calls received by a call center in a given hour. Then the range of C is
$$\text{Range}(C) = \{0, 1, 2, \ldots\},$$
which is a countable set. Therefore, C is a discrete random variable.

Remark. Understanding whether a random variable is discrete or continuous is essential for choosing the right tools to compute probabilities, expected values, and other statistical quantities.

Probability Mass Function

The probability mass function, often abbreviated as PMF, is a fundamental concept in the study of discrete random variables. It gives the probability of each possible outcome.

Definition. Let X be a discrete random variable with range $\{x_1, x_2, \ldots\}$. The **probability mass function (PMF)** of X assigns to each value x_i the probability that the random variable X takes that value, $P(X = x_i)$. The PMF must satisfy the following conditions:

- $P(X = x_i) \geq 0$ for all i,
- $\sum\limits_i P(X = x_i) = 1$.

Remark. Recall that originally we defined a probability function $\mathbb{P}$ on a sample space Ω as a function that assigns probabilities to events (subsets of Ω). If $X : \Omega \to \mathbb{R}$ is a discrete random variable, then a PMF for X can be thought of as a probability function on its range $\text{Range}(X)$.

Example. Consider a pair of six-sided dice.

[3] More generally, a random variable is a measurable function from a sample space Ω to a measurable space, often $\mathbb{R}$ or $\mathbb{R}^n$. In this course, we will restrict our attention to real-valued random variables.

1. Let Y be the random variable representing the sum of outcomes. The range of Y is $\{2, 3, \ldots, 12\}$. The PMF of Y is given by:

$$P(Y = k) = \frac{6 - |k - 7|}{36}.$$

   Moreover, the probabilities $P(Y = k)$ sum up to 1:

$$\sum_{k=2}^{12} P(Y = k) = \sum_{k=2}^{12} \frac{6 - |k - 7|}{36} = \frac{1 + 2 + 3 + 4 + 5 + 6 + 5 + 4 + 3 + 2 + 1}{36} = \frac{36}{36} = 1.$$

2. Let Z be the random variable representing the parity of the sum of outcomes, where 1 represents odd and 2 represents even. The PMF of Z is given by:

$$P(Z = i) = \begin{cases} \frac{1}{2} & \text{for } i = 1 \\ \frac{1}{2} & \text{for } i = 2. \end{cases}$$

   This reflects the equal probability of getting an odd or even sum.

Example. Baurice-Morris, a fictional individual created by Starbucks employees, goes to Las Vegas with \$20 and plays a game involving a tetrahedral die (four faces). The die is weighted, with the following probabilities for each face:

$$P(1) = 0.3, \quad P(2) = 0.2, \quad P(3) = 0.4, \quad P(4) = 0.1.$$

The outcomes affect Baurice-Morris's money as follows:

- If the die lands on 1: he wins \$5.
- If the die lands on 2: he wins nothing (\$0).
- If the die lands on 3: he loses \$7.
- If the die lands on 4: he wins \$100.

Let X be the random variable representing the total amount of money Baurice-Morris has after playing two independent rounds of the game (starting with \$20). We proceed iteratively:

After Round 1: starting with \$20, the possible outcomes and their probabilities are:

Amount	Probability
$20 + 5 = 25$	0.3
$20 + 0 = 20$	0.2
$20 - 7 = 13$	0.4
$20 + 100 = 120$	0.1

After Round 2: for each of the four outcomes above, we apply the same possible changes again using the given probabilities.

- $P(X = 6) = P(-7, -7) = P(3) \cdot P(3) = 0.4 \cdot 0.4 = 0.16$.
- $P(X = 13) = P(3, 2) + P(2, 3) = 0.4 \cdot 0.2 + 0.2 \cdot 0.4 = 0.16$.
- $P(X = 18) = P(3, 1) + P(1, 3) = 0.4 \cdot 0.3 + 0.3 \cdot 0.4 = 0.24$.
- $P(X = 20) = P(2, 2) = 0.2 \cdot 0.2 = 0.04$.
- $P(X = 25) = P(2, 1) + P(1, 2) = 0.2 \cdot 0.3 + 0.3 \cdot 0.2 = 0.12$.

- $P(X = 30) = P(1, 1) = 0.3 \cdot 0.3 = 0.09$.
- $P(X = 113) = P(3, 4) + P(4, 3) = 0.4 \cdot 0.1 + 0.1 \cdot 0.4 = 0.08$.
- $P(X = 120) = P(2, 4) + P(4, 2) = 0.2 \cdot 0.1 + 0.1 \cdot 0.2 = 0.04$.
- $P(X = 125) = P(1, 4) + P(4, 1) = 0.3 \cdot 0.1 + 0.1 \cdot 0.3 = 0.06$.
- $P(X = 220) = P(4, 4) = 0.1 \cdot 0.1 = 0.01$.

Range of X: we collect the distinct values to obtain $\{6,\ 13,\ 18,\ 20,\ 25,\ 30,\ 113,\ 120,\ 125,\ 220\}$.

The PMF of X is given by the table below:

X	$P(X)$
6	0.16
13	0.16
18	0.24
20	0.04
25	0.12
30	0.09
113	0.08
120	0.04
125	0.06
220	0.01

Check: $0.16 + 0.16 + 0.24 + 0.04 + 0.12 + 0.09 + 0.08 + 0.04 + 0.06 + 0.01 = 1$ ✓

The Beef Wellington in the newly opened Gordon Ramsay restaurant nearby costs \$100. Let's find the probability that Baurice-Morris can afford this dish after playing two rounds of the dice game. We compute the probability that the final amount X is at least \$100:

$$P(X \geq 100) = P(X = 113) + P(X = 120) + P(X = 125) + P(X = 220) = 0.08 + 0.04 + 0.06 + 0.01 = 0.19.$$

Thus, Baurice-Morris has a 19% chance of walking out of the casino with enough money to afford the Beef Wellington.

Bernoulli and Binomial Distributions

Consider a random experiment involving repeated trials with only two possible outcomes - often called *success* and *failure*. These are modeled using Bernoulli and Binomial distributions.

Let X be a random variable representing the outcome of a single trial, where success occurs with probability $p \in [0, 1]$. Then X follows a Bernoulli distribution:

$$P(X = x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

This is the basic building block of binary (yes/no) experiments, such as flipping a biased coin or checking if a lightbulb works.

If we consider n independent Bernoulli trials, each with success probability p, and let Y be the total number of successes. Then Y follows the binomial distribution with PMF

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

This distribution models the number of successes in a fixed number of independent trials. For example, the number of heads in 10 coin tosses, or the number of defective items in a sample of 20 products.

Remark. To confirm that the binomial probability mass function is valid, we must verify that the total probability sums to 1:

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

This follows directly from the binomial expansion:

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

But since $p + (1-p) = 1$, the left-hand side is equal to $1^n = 1$.

Remark. As n increases, the binomial distribution becomes more symmetric and resembles the *bell shape* of the normal distribution — especially when $p \approx 0.5$.

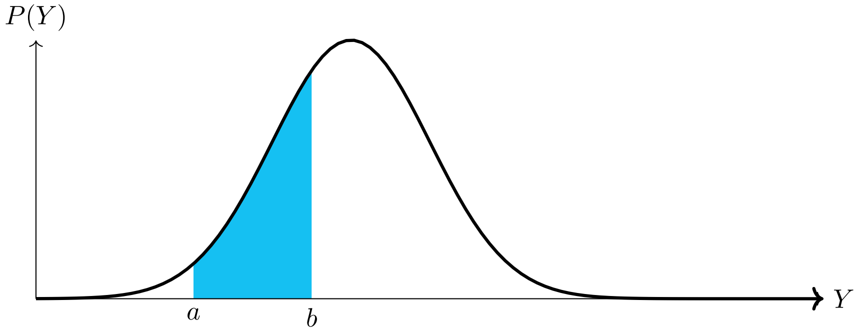
Probability Density Function

For continuous random variables, we use the probability density function (PDF) instead of the PMF. The PDF does not give the probability of a specific outcome, but rather the likelihood of the variable taking on a range of values.

Definition. For a continuous random variable Y with possible values in an interval $[a, b]$, the **probability density function (PDF)** $f_Y(y)$ is defined such that for any $a \leq y \leq b$, the probability that Y falls in the interval $[a, y]$ is given by:

$$P(a \leq Y \leq x) = \int\limits_{a}^{y} f_Y(t)\, dt$$

In essence, the PDF represents the density of probabilities along the range of Y.



The shaded region represents the probability $P(a \leq Y \leq b)$ for a continuous random variable Y with values in the interval $[a, b]$. This probability is calculated by finding the area under the probability density function $f_Y(y)$ between a and b. In mathematical terms, it is given by the definite integral $\int\limits_{a}^{b} f_Y(t)\, dt$.

The PDF is a continuous analog to the PMF, suited for variables that can take on infinitely many values. In order for a function $f : \Omega \to \mathbb{R}$ to be a valid probability density function, it must satisfy the following properties:

1. $f(t) \geq 0$ for all t in the sample space Ω. Otherwise, when we integrate, we could get a negative number as our probability.

2. $\int\limits_{-\infty}^{\infty} f(t)\, dt = 1$. This corresponds to the entire sample space having probability 1.

These properties ensure that the PDF properly describes the likelihood of the random variable taking on different values within the sample space.

Example. 1. Let

$$f(t) = \begin{cases} 2t & \text{if } 0 \leq t \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Question. Is this a valid PDF?

We verify the two required conditions:

- $f(t) \geq 0$ for all $t \in \mathbb{R}$.
- $\int\limits_{-\infty}^{\infty} f(t)\,dt = \int\limits_{0}^{1} 2t\,dt = t^2 \Big|_0^1 = 1 - 0 = 1.$

Therefore, $f(t)$ is a valid PDF.

2. Let

$$f(x) = \begin{cases} x^2 & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Question. Is this a valid PDF?

We need to check two things. First, that $f(x)$ is non-negative (which it is), and second that it integrates out to 1. We have

$$\int\limits_{-\infty}^{\infty} f(x)\,dx = \int\limits_{0}^{2} x^2\,dx = \frac{x^3}{3}\Big|_0^2 = \frac{8}{3} \neq 1.$$

This function does not satisfy the properties required for a valid probability density function.

3. Let

$$f(x) = \begin{cases} \ell n(2) \cdot 2^{1-x} & \text{if } x > 1 \\ 0 & \text{otherwise.} \end{cases}$$

Question. Is this a valid PDF?

First, $f(x)$ is non-negative. Next, we icompute:

$$\int\limits_{1}^{\infty} \ell n(2) 2^{1-x}\,dx = -\frac{\ell n(2)}{\ell n(2)} \cdot 2^{1-x}\Big|_1^{\infty} = -2^{-x+1}\Big|_1^{\infty} = 0 + 1 = 1.$$

So yes, this is a valid PDF.

Fun fact: in both mathematics and document formatting, "PDF" is important. In probability, it stands for *probability density function*, while in tech, it means *portable document format.* One describes how probability spreads, the other how documents do!

Uniform Distribution

The **uniform distribution** is a fundamental example of a continuous distribution where all values in a certain interval are equally likely.

Definition. A random variable X is said to be **uniformly distributed** on the interval $[a, b]$ if its PDF is constant over that interval:

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

We check the required properties:

- $f_X(x) \ge 0$;
- $\int\limits_{-\infty}^{\infty} f_X(x)\,dx = \int\limits_{a}^{b} \frac{1}{b-a}\,dx = \frac{x}{b-a}\Big|_a^b = \frac{b-a}{b-a} = 1.$

So this is indeed a valid PDF.

Lecture 8
Cumulative Distribution Function

In probability theory, it is often important to understand the likelihood that a random variable takes on a value less than or equal to a certain number. This leads us to the concept of the *cumulative distribution function*, which captures the entire probability distribution of a random variable up to any given threshold.

Definition. The **cumulative distribution function** (CDF) of a random variable X, denoted by $F_X(x)$, is defined for each $x \in \mathbb{R}$ by

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

That is, $F_X(x)$ gives the probability that the random variable X takes on a value less than or equal to x.

In other words, it gives the probability that the random variable X is less than or equal to x.

Relationship Between the PDF and CDF

Let X be a continuous random variable with probability density function (PDF) $f_X(x)$. The cumulative distribution function (CDF) $F_X(x)$ is defined as the probability that X takes a value less than or equal to x:

$$F_X(x) = P(X \le x) = \int\limits_{-\infty}^{x} f_X(t)\,dt.$$

Conversely, if $F_X(x)$ is differentiable at x, then the PDF can be recovered by differentiation:

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

Analogy with Mass and Density

The relationship between the PDF and CDF is analogous to the relationship between density and mass in physics, particularly in one-dimensional settings.

- In physics, suppose we have a thin rod lying along a line, with mass distributed non-uniformly. The linear mass density $\rho(t)$ describes how mass is distributed at point t along the rod. The total mass accumulated up to position x is given by:

$$M(x) = \int\limits_{-\infty}^{x} \rho(t)\,dt.$$

- Similarly, in probability theory, the PDF $f_X(x)$ measures the "concentration" of probability around x. The total probability accumulated up to point x is given by the corresponding value of the CDF:

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

So just as mass accumulates along the rod according to the density $\rho(t)$, probability accumulates along the real line according to the density $f_X(t)$. The CDF plays the role of cumulative mass: it tells us the total probability up to a given point.

Example. 1. Given the PDF $f_X(x) = \frac{1}{3}$ for $0 \le x \le 3$ and $f_X(x) = 0$ outside of this interval, find the probability that X is less than or equal to 2, i.e., $P(X \le 2)$.

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt = \int_{0}^{x} \frac{1}{3}\,dt = \frac{x}{3}.$$

So, $P(X \le 2) = F_X(2) = \frac{2}{3}$.

2. Given the PDF $f_X(x) = 2x$ for $0 \le x \le 1$, find the probability that X is greater than 0.5, i.e., $P(X > 0.5)$.

$$P(X > 0.5) = 1 - P(X \le 0.5) = 1 - F_X(0.5) = 1 - \int_{0}^{0.5} 2t\,dt = 1 - t^2\Big|_{0}^{0.5} = 1 - 0.25 = 0.75.$$

Some Important CDFs

We give examples of CDFs for some important continuous distributions.

Example. 1. **Continuous Uniform Distribution on** $[a, b]$.

The PDF is:

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

The CDF is computed by integrating the PDF:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } a \le x < b, \\ 1 & \text{if } x \ge b. \end{cases}$$

2. **Exponential Distribution with parameter** $\lambda > 0$.

The PDF is:

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

We compute the CDF by integrating the PDF:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt = \begin{cases} 0 & \text{if } x < 0, \\ \int_{0}^{x} \lambda e^{-\lambda t}\,dt = -e^{-\lambda t}\Big|_{0}^{x} = 1 - e^{-\lambda x} & \text{if } x \ge 0. \end{cases}$$

Remark. This distribution is widely used to model waiting times between independent events in a Poisson process, such as time between arrival of customers.

3. **Standard Normal Distribution**

   The PDF is:
   $$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

   The CDF is given by:
   $$F_X(x) = \int\limits_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = \Phi(x).$$

   This integral cannot be expressed in closed form using elementary functions and is typically evaluated numerically or looked up in standard normal tables.

   Remark. The normal distribution is extremely important in probability and statistics because it often appears when we look at the combined effect of many small, random influences. Even when individual factors are not normally distributed, their sum or average often behaves like a normal distribution (due to the *Central Limit Theorem*).

Properties of the CDF

Let $F_X(x) = P(X \leq x)$ be the cumulative distribution function of a random variable X. The CDF has the following fundamental properties:

1. **Non-decreasing:** the CDF is a non-decreasing function. That is, for any $a \leq b$, we have
   $$F_X(a) \leq F_X(b).$$
   This follows from the definition of the CDF, since the event $\{X \leq a\}$ is a subset of the event $\{X \leq b\}$ whenever $a \leq b$.

2. **Limits at Infinity:** the CDF satisfies
   $$\lim_{x \to -\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F_X(x) = 1.$$
   These limits reflect the fact that the probability of observing values far below or far above the range of the random variable converges to 0 and 1, respectively.

3. **Bounded Between 0 and 1:** for all $x \in \mathbb{R}$, the CDF satisfies
   $$0 \leq F_X(x) \leq 1.$$
   This is a direct consequence of the definition of $F_X(x)$ as a probability function.

Lecture 9

Transformation of Random Variables

In probability theory, we often encounter situations where we need to analyze the behavior of a new random variable that is derived from one or more existing random variables. This process is known as the transformation of random variables. It allows us to model complex phenomena and make predictions about the outcomes.

The transformation of random variables is, essentially, a composition of functions. Consider a sample space Ω and random variables X and Y, where $X : \Omega \to \mathbb{R}$ and $Y : \mathbb{R} \to \mathbb{R}$. The random variable Y is a *transformation* of X if there exists a function $g : \mathbb{R} \to \mathbb{R}$ such that $Y = g(X)$.

This concept enables us to analyze the relationship between the original and transformed random variables. The properties of the transformation function g play a crucial role in understanding the resulting probability distribution.

The computation of the resulting distribution may be straightforward. For instance, if g is a linear function. However, in more complex cases, numerical methods or specialized techniques may be required to analyze the transformed random variable.

By understanding the properties of the transformation function, we can analyze and make predictions about complex phenomena.

Discrete Random Variables

Example. Suppose we are analyzing the customer ratings of products on an e-commerce platform. Let X be the rating of a product on a scale of 1 to 5. The PMF of X can be defined based on the customer ratings:

$$P_X(x) = \begin{cases} 0.1 & \text{if } x = 1 \\ 0.2 & \text{if } x = 2 \\ 0.4 & \text{if } x = 3 \\ 0.2 & \text{if } x = 4 \\ 0.1 & \text{if } x = 5. \end{cases}$$

Now, let's consider the transformation $Y = 6 - X$, which represents the inverse of the rating scale. This transformation is natural because it measures the level of dissatisfaction (higher rating means lower dissatisfaction).

To compute the PMF of Y, we consider all possible values of Y (ranging from 1 to 5) and compute their respective probabilities.

For instance, $P(Y = 1)$ is the probability that a customer is extremely dissatisfied, which is the same as $P(6-X = 1) = P(X = 5) = 0.1$, while $P(Y = 2)$ is the probability of a relatively low level of dissatisfaction, which is the same as $P(X = 4) = 0.2$, etc.

This transformation provides insights into the level of satisfaction expressed by customers.

Example. Suppose we are measuring the fluctuations from a certain standard in a physical process. Let X be the fluctuation, which can take values -2, -1, 0, 1, 2, and 3. The PMF of X is given by:

$$P_X(x) = \begin{cases} 0.1 & \text{if } x = -2 \\ 0.2 & \text{if } x = -1 \\ 0.3 & \text{if } x = 0 \\ 0.2 & \text{if } x = 1 \\ 0.1 & \text{if } x = 2 \\ 0.1 & \text{if } x = 3 \\ 0 & \text{otherwise.} \end{cases}$$

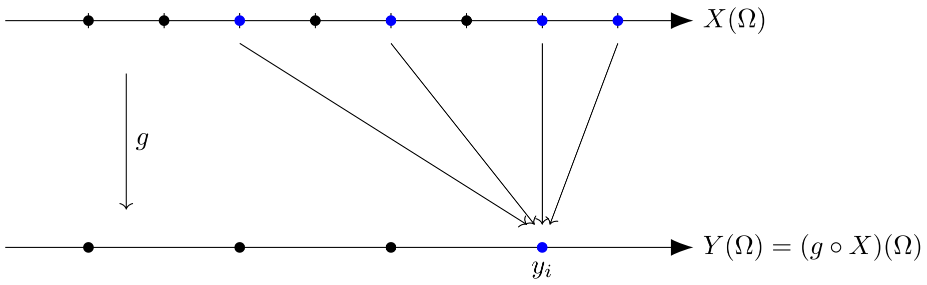
Now, let's consider the transformation $Y = X^2$, which represents the squared fluctuations. This transformation is natural because it emphasizes the magnitude of the fluctuations, regardless of their direction. The range of Y is $\{0, 1, 4, 9\}$. To compute the PMF of Y, we consider each possible value:

$$\begin{aligned} P_Y(0) &= P(X^2 = 0) = P(X = 0) = 0.3 \\ P_Y(1) &= P(X^2 = 1) = P(X = -1 \text{ or } X = 1) = 0.2 + 0.2 = 0.4 \\ P_Y(4) &= P(X^2 = 4) = P(X = -2 \text{ or } X = 2) = 0.1 + 0.1 = 0.2 \\ P_Y(9) &= P(X^2 = 9) = P(X = -3 \text{ or } X = 3) = 0.1 + 0.1 = 0.2. \end{aligned}$$

For all other values of y, $P_Y(y) = 0$.

Summary

When a new discrete random variable $Y = g(X)$ is defined via a function of an existing random variable X, its PMF is obtained by summing the probabilities of all values of X that map to each y_i. This is often called the *pushforward distribution* or computing the PMF via the *preimage rule*:



$$P(Y = y_i) = \sum_{x, g(x) = y_i} P(X = x)$$

Continuous Random Variables

Example. Consider a continuous random variable X with the probability density function:

$$f_X(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

This is a valid PDF, as it is nonnegative and integrates to 1 over its support:

$$\int_0^1 2x \, dx = x^2 \Big|_0^1 = 1.$$

Now define a new random variable $Y = X^2$. We aim to compute the CDF $F_Y(y) = P(Y \leq y)$. To find $F_Y(y)$, we compute:

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = P(X \leq \sqrt{y}),$$

since $X \geq 0$. Therefore, the CDF of $Y = X^2$ is:

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0, \\ \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x) \, dx = \int_0^{\sqrt{y}} 2x \, dx = x^2 \Big|_0^{\sqrt{y}} = y & \text{if } 0 \leq y \leq 1, \\ 1 & \text{if } y > 1. \end{cases}$$

Remark. Notice that the PDF of Y can be obtained directly from the definition of a transformed continuous random variable. Since $Y = g(X) = X^2$ and $X \in [0, 1]$, the inverse of the transformation is $g^{-1}(y) = \sqrt{y}$, which is valid for $y \in [0, 1]$.

Using the change-of-variable formula we can find that the PDF of Y is given by:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left|(g^{-1}(y))'\right| = f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}}.$$

Since $f_X(x) = 2x$, we have:

$$f_Y(y) = \begin{cases} 2\sqrt{y} \cdot \frac{1}{2\sqrt{y}} = 1 & \text{if } 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Alternatively, we can find $f_Y(y)$ by differentiating the CDF $F_Y(y) = y$:

$$f_Y(y) = \frac{d}{dy}(F_Y(y)) = \begin{cases} \frac{d(y)}{dy} = 1 & \text{if } 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let's take a look at more examples.

Linear Transformation

Let X be a continuous random variable uniformly distributed on the interval $[-1, 2]$. Consider the transformation $Y = 2X + 3$. Our goal is to find the cumulative distribution function and probability density function of Y.

We begin by expressing the CDF of Y in terms of X:

$$F_Y(y) = P(Y \leq y) = P(2X + 3 \leq y) = P\left(X \leq \frac{y-3}{2}\right) = F_X\left(\frac{y-3}{2}\right).$$

The CDF of a uniform random variable on the interval $[-1, 2]$ is given by:

$$F_X(x) = \begin{cases} 0 & \text{if } x < -1, \\ \frac{x+1}{3} & \text{if } -1 \leq x \leq 2, \\ 1 & \text{if } x > 2. \end{cases}$$

Now, substitute $x = \frac{y-3}{2}$ into this expression to get the CDF of Y:

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 1, \\ \frac{y-1}{6} & \text{if } 1 \leq y \leq 7, \\ 1 & \text{if } y > 7. \end{cases}$$

To obtain the PDF $f_Y(y)$, we differentiate the CDF:

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \begin{cases} \frac{1}{6} & \text{if } 1 < y < 7, \\ 0 & \text{otherwise.} \end{cases}$$

So, the transformed variable $Y = 2X+3$ follows a uniform distribution on the interval $[1, 7]$, with constant density $\frac{1}{6}$.

Logarithmic Transformation

Suppose X is an exponential random variable with rate parameter $\lambda = 3$. That is, the probability density function of X is given by:

$$f_X(x) = \begin{cases} 3e^{-3x} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

We are interested in the transformed random variable $Y = \ell n(X)$. To find the distribution of Y, we start by computing its cumulative distribution function:

$$F_Y(y) = P(Y \leq y) = P(\ell n(X) \leq y) = P(X \leq e^y).$$

Since $F_X(x) = P(X \leq x)$, we substitute $x = e^y$ into the CDF of the exponential distribution. First, recall from the previous lecture that the CDF of X is:

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - e^{-3x} & \text{if } x \geq 0. \end{cases}$$

Now plug in $x = e^y$:

$$F_Y(y) = F_X(e^y) = 1 - e^{-3e^y}.$$

Notice that $e^y > 0$ for all y, so the cumulative distribution function of $Y = \ell n(X)$ is:

$$F_Y(y) = 1 - e^{-3e^y}.$$

Lecture 10
Expected Value

We are all familiar with the arithmetic mean - the usual way to compute the average of a list of numbers. However, in probability theory, outcomes often occur with different likelihoods. In such situations, a simple arithmetic mean does not accurately capture the overall behavior.

Instead, we compute a *weighted average*, where each value is multiplied by the probability of its occurrence. This leads us to the concept of the *expected value* of a random variable.

Example. Consider a biased six-sided die, where the probabilities of rolling each face are given by:

$$\begin{aligned} &P(X=1)=0.1, \qquad &&P(X=2)=0.2, \quad &&P(X=3)=0.3, \\ &P(X=4)=0.15, \qquad &&P(X=5)=0.1, \quad &&P(X=6)=0.15. \end{aligned}$$

The expected value of X, denoted $\mathbb{E}(X)$, is computed as a weighted average of the outcomes:

$$\begin{aligned} \mathbb{E}(X) &= (1\cdot 0.1)+(2\cdot 0.2)+(3\cdot 0.3)+(4\cdot 0.15)+(5\cdot 0.1)+(6\cdot 0.15)= \\ &= 0.1+0.4+0.9+0.6+0.5+0.9=3.4. \end{aligned}$$

So, for this biased die, the expected value is $\mathbb{E}(X)=3.4$.

For comparison, consider a fair six-sided die, where each face has an equal probability of $\frac{1}{6}$. The expected value is the arithmetic mean of all outcomes:

$$\mathbb{E}(X)=\frac{1+2+3+4+5+6}{6}=\frac{21}{6}=3.5.$$

Conclusion. The expected value for the biased die is slightly lower than that of the fair die. This illustrates how unequal probabilities (or “bias”) shift the expected value from the uniform mean.

Definition. The **expected value** of a random variable X, denoted as $\mathbb{E}(X)$ or μ_X, is defined as

$$\mathbb{E}(X)=\sum_x x\cdot P(X=x)$$

if X is discrete. For a continuous random variable, it is defined as

$$\mathbb{E}(X)=\int\limits_{-\infty}^{\infty} x\cdot f_X(x)\,dx.$$

The difference lies in the summation for finite/discrete sample spaces and integration for continuous random variables.

Expected Value as a Center of Mass (Optional)

This section is optional and intended to provide physical intuition for the concept of expected value, by drawing an analogy with the center of mass in physics.

Before, we saw that the cumulative distribution function is analogous to mass. Similarly, the expected value is the probabilistic analogue of the *center of mass.*

Discrete case: for a discrete random variable X, you can imagine placing a point mass at each value x on the number line, with weight $P(X=x)$. The expected value is then the “balance point” of this system — much like the center of mass of a set of discrete weights:

$$\mathbb{E}(X)=\sum_x x\cdot P(X=x).$$

Continuous case: for a continuous random variable, the analogy continues.In one-dimensional physics, if a rod has mass density $\rho(x)$ along a line, then its center of mass is given by

$$\frac{\int x\rho(x)\,dx}{\int \rho(x)\,dx}.$$

In probability theory, the density function $f_X(x)$ plays the role of a *normalized* mass distribution (since $\int\limits_{-\infty}^{\infty} f_X(x)\,dx = 1$), so the expected value becomes:

$$\mathbb{E}(X) = \int\limits_{-\infty}^{\infty} x \cdot f_X(x)\,dx.$$

Thus, the expected value of a random variable can be intuitively interpreted as its probabilistic "center of mass" — the average location where the distribution would balance if it were a physical object.

Discrete Sample Spaces

Biased Coin Toss, I Suppose we have a biased coin with probabilities: $P(H) = 0.4$ and $P(T) = 0.6$. We define a random variable X as follows:

$$X(H) = 1, \quad X(T) = 2.$$

The expected value is calculated as:

$$\mathbb{E}(X) = P(H) \cdot 1 + P(T) \cdot 2 = 1 \cdot 0.4 + 2 \cdot 0.6 = 1.6.$$

Biased Coin Toss, II Suppose we have a biased coin with probabilities $P(H) = 0.4$ and $P(T) = 0.6$. We want to compute the expected number of tosses needed to get the first Heads.

Let N be the number of tosses before the first Heads appears. The expected value of N, denoted $\mathbb{E}(N)$, is given by summing over the possible outcomes:

$$\mathbb{E}(N) = P(H) \cdot 0 + P(T)P(H) \cdot 1 + P(T)^2 P(H) \cdot 2 + \cdots .$$

This is a weighted average over the number of failed tosses (Tails) before the first success (Heads). Let us set $t = P(T)$, and write:

$$\mathbb{E}(N) = P(H) \sum_{n=0}^{\infty} n t^n.$$

We now define a *generating function*:

$$G(t) = \sum_{n=0}^{\infty} t^n = \frac{1}{1-t},$$

so that

$$\sum_{n=0}^{\infty} n t^n = t \cdot G'(t) = \frac{t}{(1-t)^2}.$$

Substituting back, we obtain:

$$\mathbb{E}(N) = P(H) \cdot \frac{t}{(1-t)^2} = P(H) \cdot \frac{P(T)}{(1-P(T))^2}.$$

Plugging in the numbers:

$$\mathbb{E}(N) = 0.4 \cdot \frac{0.6}{(1-0.6)^2} = 0.4 \cdot \frac{0.6}{0.16} = 1.5.$$

Question. Notice that the final answer simplifies to $\mathbb{E}(N) = \dfrac{P(T)}{P(H)}$. Is there a quicker way to deduce this?

Yes 🙂
We can model the expected value recursively by conditioning on the outcome of the first toss.

$$\mathbb{E}(N) = P(H) \cdot 0 + P(T) \cdot (\mathbb{E}(N) + 1).$$

Solving the equation:

$$\begin{aligned}
&\mathbb{E}(N) = 0.6 \cdot (\mathbb{E}(N) + 1) = 0.6\mathbb{E}(N) + 0.6 \\
&\mathbb{E}(N) - 0.6\mathbb{E}(N) = 0.6 \\
&0.4\mathbb{E}(N) = 0.6 \\
&\mathbb{E}(N) = \frac{0.6}{0.4} = 1.5
\end{aligned}$$

This method confirms the result and provides an intuitive way of thinking about the problem: after each Tail, we effectively ”restart” the process and add 1 to the expected count.

Generating Functions (Optional)

This section is optional and intended for students interested in a deeper exploration of discrete structures and algebraic tools. Generating functions are a powerful tool that transform sequences into power series, enabling us to analyze and solve problems involving recurrence relations, sums, and expected values more effectively.

We have already seen an example of this technique when computing the expected number of tosses until the first Heads appears for a biased coin. Let us now explore this method more formally through two classic examples.

Example.

1. **Geometric Series.** Consider the geometric series $S(t) = 1 + t + t^2 + t^3 + \cdots$. To find a closed-form expression for this series, we can manipulate it as follows:

$$tS(t) = t + t^2 + t^3 + \cdots$$

Subtracting the two series:

$$S(t) - tS(t) = 1 \quad \Rightarrow \quad S(t)(1-t) = 1 \quad \Rightarrow \quad S(t) = \frac{1}{1-t}.$$

2. **Fibonacci Numbers.** Let F_n denote the Fibonacci sequence defined by $F_0 = 0$, $F_1 = 1$, and recursively as $F_{n+2} = F_{n+1} + F_n$. We can encode the Fibonacci sequence in a generating function:

$$F(t) = \sum_{n=0}^{\infty} F_n t^n = F_0 + F_1 t + F_2 t^2 + F_3 t^3 + \cdots$$

Next, let us compute the shifted versions of the generating function. Multiplying a generating function by t has the effect of shifting the entire sequence one step to the right (increasing each power of t by 1). Likewise, multiplying by t^2 shifts it two steps:

$$\begin{aligned}
tF(t) &= F_0 t + F_1 t^2 + F_2 t^3 + F_3 t^4 + \cdots \\
t^2 F(t) &= F_0 t^2 + F_1 t^3 + F_2 t^4 + F_3 t^5 + \cdots
\end{aligned}$$

Now subtract both of these from the original generating function:

$$F(t) - tF(t) - t^2F(t) = F_0 + (F_1 - F_0)t + (F_2 - F_1 - F_0)t^2 + (F_3 - F_2 - F_1)t^3 + (F_4 - F_3 - F_2)t^4 + \cdots$$

This gives the identity:

$$F(t) - tF(t) - t^2F(t) = F_0 + (F_1 - F_0)t + (F_2 - F_1 - F_0)t^2 + \cdots + (F_{n+2} - F_{n+1} - F_n)t^{n+2} + \cdots$$

Since the Fibonacci sequence satisfies the recurrence $F_{n+2} = F_{n+1} + F_n$, all coefficients from t^2 onward vanish. Substituting the initial values $F_0 = 0$, $F_1 = 1$, we obtain:

$$F(t) - tF(t) - t^2F(t) = t.$$

Solving for $F(t)$, we conclude:

$$F(t) = \frac{t}{1 - t - t^2}.$$

Remark. Multiplying a generating function by t shifts its coefficients to higher degrees, effectively *delaying* the sequence. In contrast, taking the derivative of a generating function has the opposite effect: it multiplies each coefficient by its index and shifts the sequence leftward (removing the constant term). This reflects the relationship:

$$F(t) = \sum_{n=0}^{\infty} a_n t^n \quad \Rightarrow \quad F'(t) = \sum_{n=1}^{\infty} n a_n t^{n-1}.$$

This derivative trick appeared in our earlier example with the biased coin.

Continuous Random Variables

- **Uniform Distribution on $[0, 1]$**
  Let X have a uniform distribution on the interval $[0, 1]$. The probability density function is $f_X(x) = 1$ for $0 \leq x \leq 1$. The expected value is computed as:

  $$\mathbb{E}(X) = \int_0^1 x \cdot 1 \, dx = \left. \frac{x^2}{2} \right|_0^1 = \frac{1}{2}.$$

- **Uniform Distribution on $[-2, 2]$**
  Let W have a uniform distribution on the interval $[-2, 2]$, with PDF $f_W(w) = \frac{1}{4}$ for $-2 \leq w \leq 2$. The expected value is:

  $$\mathbb{E}(W) = \int_{-2}^{2} w \cdot \frac{1}{4} \, dw = \frac{1}{4} \int_{-2}^{2} w \, dw = \frac{1}{4} \left. \frac{w^2}{2} \right|_{-2}^{2} = \frac{1}{4} \cdot 0 = 0.$$

  Remark. This result aligns with the center-of-mass interpretation of expected value. Imagine the uniform distribution as a thin, evenly dense rod lying along the number line from -2 to 2. If you were to try to balance this rod on the tip of your finger, you would naturally place your finger at the midpoint — zero — to keep it balanced. That point of balance corresponds exactly to the expected value.

- **Exponential Distribution**
  Let Y be an exponential random variable with parameter $\lambda > 0$, so that $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$. The expected value is:

  $$\mathbb{E}(Y) = \int_0^{\infty} y \cdot \lambda e^{-\lambda y} \, dy.$$

  We evaluate this using integration by parts. Let $u = y$, so $du = dy$, and let $dv = \lambda e^{-\lambda y} dy$, so $v = -e^{-\lambda y}$. Then:

  $$\mathbb{E}(Y) = \left. -ye^{-\lambda y} \right|_0^{\infty} + \int_0^{\infty} e^{-\lambda y} \, dy.$$

  The first term vanishes at both ends:

  $$\lim_{y \to \infty} ye^{-\lambda y} = 0, \quad \left. ye^{-\lambda y} \right|_{y=0} = 0.$$

The remaining integral is:

$$\int_0^\infty e^{-\lambda y}\,dy = \frac{-1}{\lambda}e^{-\lambda y}\Big|_0^\infty = \frac{1}{\lambda}.$$

Hence, the expected value is:

$$\mathbb{E}(Y) = \frac{1}{\lambda}.$$

Properties of Expectation

- $\mathbb{E}(c) = c$: the expected value of a constant random variable is the constant itself.
- Linearity: for any constants a and b, and random variables X and Y,

  $$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

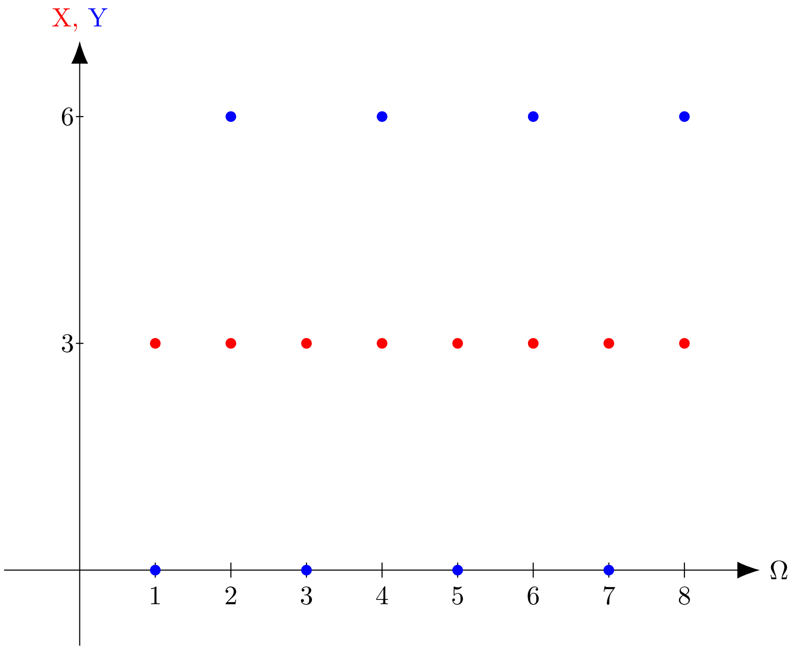
  This property states that the expectation of a linear combination of random variables is equal to the linear combination of their individual expectations.

Joke

Why did the statistician go broke? Because he had too many expectations from random variables!

Lecture 11
Variance and Standard Deviation

In statistics, variance and standard deviation are crucial measures that help us understand the spread or dispersion of a random variable's outcomes. While expectation gives us a measure of central tendency, variance and standard deviation provide us with information about how spread out the values are. Consider two random variables X and Y uniformly distributed on sample space $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8\}$ with the same mean $\mathbb{E}(X) = \mathbb{E}(Y) = 3$. The first one, X, has all its values equal to that mean, while Y has a much more spread out distribution:



Definition. The **variance** of a random variable X, denoted by $\mathrm{Var}(X)$, measures the expected squared deviation from its mean. It is defined as

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

We now derive an equivalent, often more convenient, formula for computing variance.

Let $\mu_X := \mathbb{E}(X)$. Then, using properties of expectation, we compute:

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mu_X)^2\right] = \mathbb{E}\left[X^2 - 2\mu_X X + \mu_X^2\right] = \mathbb{E}(X^2) - 2\mu_X \mathbb{E}(X) + \mathbb{E}(\mu_X^2) =$$
$$\mathbb{E}(X^2) - 2\mu_X^2 + \mu_X^2 = \mathbb{E}(X^2) - \mu_X^2.$$

Thus, we obtain the alternate formula:

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

This formula is often more convenient for practical calculations.

Definition. The **standard deviation** of a random variable X, denoted as σ_X, is the square root of the variance

$$\sigma_X = \sqrt{\mathrm{Var}(X)}.$$

Finite Sample Spaces

1. **Coin Toss.** Consider the biased coin toss with probabilities: $P(H) = 0.4$ and $P(T) = 0.6$. Let X be the random variable defined as $X(H) = 1$ and $X(T) = 2$. The expected value is $\mathbb{E}(X) = 1.6$. The variance is calculated as:

   $$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = (1 - 1.6)^2 \cdot 0.4 + (2 - 1.6)^2 \cdot 0.6 = 0.24.$$

   The standard deviation is then $\sigma_X = \sqrt{\mathrm{Var}(X)} \approx 0.49$.

2. **Biased Die.** Consider the biased six-sided die with probabilities:

   $$P(X = 1) = 0.1,\ P(X = 2) = 0.2,\ P(X = 3) = 0.3,$$

   $$P(X = 4) = 0.15,\ P(X = 5) = 0.1,\ P(X = 6) = 0.15.$$

   The expected value is $\mathbb{E}(X) = 3.4$. The variance is computed as:

   $$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = (1 - 3.4)^2 \cdot 0.1 + (2 - 3.4)^2 \cdot 0.2 + \ldots + (6 - 3.4)^2 \cdot 0.15 = 2.34$$

   The standard deviation is then $\sigma_X = \sqrt{\mathrm{Var}(X)} \approx 1.53$. Let's use the alternative formula for variance to verify that it indeed yields the same result:

   $$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = (1^2 \cdot 0.1) + (2^2 \cdot 0.2) + \ldots + (6^2 \cdot 0.15) - (3.4)^2 = 2.34.$$

   As expected, the result coincides with the previous calculation using the original formula. This demonstrates that both formulas for variance are equivalent and yield the same result.

Continuous Random Variables

For a continuous random variable X with probability density function $f_X(x)$, the variance is defined as

$$\mathrm{Var}(X) = \int\limits_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 \cdot f_X(x)\, dx = \int\limits_{-\infty}^{\infty} x^2 \cdot f_X(x)\, dx - \mathbb{E}^2(X).$$

1. **Uniform Distribution on** $[0, 1]$**.** Let X have a uniform distribution on the interval $[0, 1]$. The PDF is $f_X(x) = 1$ for $0 \le x \le 1$. The expected value is $\mathbb{E}(X) = \frac{1}{2}$. The variance is computed as

$$\mathrm{Var}(X) = \int\limits_0^1 x^2\,dx - \frac{1}{4} = \frac{x^3}{3}\bigg|_0^1 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

The standard deviation is then $\sigma_X = \sqrt{\mathrm{Var}(X)} \approx 0.29$.

2. **Exponential Distribution.** Consider a random variable Y with an exponential distribution, $f_Y(y) = \lambda e^{-\lambda y}$ for $y > 0$. Recall (from the previous lecture) that the expected value is

$$\mathbb{E}(Y) = \int\limits_0^\infty y \cdot \lambda e^{-\lambda y}\,dy = -ye^{-\lambda y}\bigg|_0^\infty + \int\limits_0^\infty e^{-\lambda y}\,dy = \int\limits_0^\infty e^{-\lambda y}\,dy = -\frac{e^{-\lambda y}}{\lambda}\bigg|_0^\infty = \frac{1}{\lambda},$$

where we have used integration by parts.

Using integration by parts and the outcome of the above calculation, we find the variance as

$$\mathrm{Var}(Y) = \int\limits_0^\infty y^2 \cdot \lambda e^{-\lambda y}\,dy - \frac{1}{\lambda^2} = -y^2 e^{-\lambda y}\bigg|_0^\infty + \int\limits_0^\infty 2ye^{-\lambda y}\,dy - \frac{1}{\lambda^2} = \frac{2}{\lambda}\int\limits_0^\infty y \cdot \lambda e^{-\lambda y}\,dy - \frac{1}{\lambda^2} = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

The standard deviation is then $\sigma_Y = \sqrt{\mathrm{Var}(Y)} = \frac{1}{\lambda}$.

3. Let Z have the probability density function:

$$f_Z(z) = 0.1(3z^2 + 1) \text{ for } 0 \le z \le 2.$$

To verify if this is a valid PDF, we need to check:

$$\int\limits_{-\infty}^{\infty} f_Z(z)\,dz = \int\limits_0^2 0.1(3z^2 + 1)\,dz = 0.1(z^3 + z)\bigg|_0^2 = 0.1(10 - 0) = 1 \quad \checkmark$$

The expected value is:

$$\mathbb{E}(Z) = \int\limits_0^2 z \cdot 0.1(3z^2 + 1)\,dz = 0.1\int\limits_0^2 (3z^3 + z)\,dz = 0.1\left(\frac{3}{4}z^4 + \frac{1}{2}z^2\right)\bigg|_0^2 = 0.1 \cdot (12 + 2) = 1.4.$$

The variance is computed as

$$\mathrm{Var}(Z) = \int\limits_0^2 z^2 \cdot 0.1(3z^2+1)\,dz - (1.4)^2 = 0.1\int\limits_0^2 3z^4 + z^2\,dz - 1.96 = 0.1\left(\frac{3z^5}{5} + \frac{z^3}{3}\right)\bigg|_0^2 = 0.1\left(\frac{96}{5} + \frac{8}{3}\right) \approx 2.19.$$

The standard deviation is then $\sigma_Z \approx \sqrt{2.19} \approx 1.48$.

Properties of Variance

- For any constant c, $\mathrm{Var}(c) = 0$. A constant random variable has zero variance.
- For any constant a and random variable X,

$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X).$$

Scaling a random variable by a constant scales the variance by the square of that constant.

- If X and Y are independent,
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$
This does not hold in general without independence.

Lecture 12
Markov's and Chebyshev's Inequalities

In probability theory, inequalities like Markov's and Chebyshev's provide powerful tools for estimating how likely a random variable is to deviate from its expected value — even when we lack full knowledge of its distribution.

Markov's Inequality gives an upper bound on the probability that a non-negative random variable exceeds a given multiple of its expectation. It is especially useful when only the expected value is known.

Chebyshev's Inequality goes further by incorporating information about the variance, yielding tighter bounds on how far a random variable can deviate from its mean.

Markov's Inequality

Consider a non-negative random variable X, and let $a > 0$ be any positive number. **Markov's Inequality** states:
$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$
In words, this inequality bounds the probability that X exceeds the threshold a by the ratio of its expected value to a. It holds for *any* non-negative random variable, regardless of its distribution.

Example. 1. Suppose we have a biased coin with $P(H) = 0.2$ and we flip it ten times. We want to estimate the probability of obtaining at least eight Heads.

Let X be the random variable representing the number of Heads obtained in ten flips. The expected value of X is $\mathbb{E}(X) = n \cdot p = 10 \cdot 0.2 = 2$. Using Markov's inequality with $a = 8$, we get
$$P(X \geq 8) \leq \frac{2}{8} = 0.25.$$
Now, let's compute the actual probability:
$$P(X \geq 8) = P(X=8)+P(X=9)+P(X=10) = \binom{10}{8}\cdot 0.2^8 \cdot 0.8^2 + \binom{10}{9}\cdot 0.2^9 \cdot 0.8 + 0.2^{10} \approx 0.00007793.$$
Conclusion: the inequality $P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$ gives an upper bound for the probability. In this case, the upper bound was 0.25, which is much larger than the actual probability of approximately 0.00007793. While this bound is often not very tight, it is a quick and easy way to obtain an estimate for the probability of rare events using the expected value.

2. In a company with 200 employees, it is found that the average number of years each employee has worked at the company is 6. Using Markov's inequality, we can find
$$P(\text{Years of Service} \geq 10) \leq \frac{\mathbb{E}(\text{Years of Service})}{10} = \frac{6}{10} = 0.6.$$
This means that the probability of an employee having worked at the company for 10 or more years is less than or equal to 0.6.

3. In a city with a population of 100000, you know that the average number of cars per household is 1.5. Using Markov's inequality, we can find

$$P(\text{Number of Cars in a Household} \geq 3) \leq \frac{\mathbb{E}(\text{Number of Cars in a Household})}{3} = \frac{1.5}{3} = 0.5$$

   This means that the probability of a household having 3 or more cars is less than or equal to 0.5.

4. Let X be a non-negative continuous random variable with probability density function given by

$$f(x) = \begin{cases} \frac{x}{8} & \text{if } 0 \leq x \leq 4 \\ 0 & \text{otherwise.} \end{cases}$$

   .

   First we find $\mathbb{E}(X) = \int\limits_0^4 \frac{x^2}{8} = \frac{x^3}{24}\Big|_0^4 = \frac{8}{3}$. Using Markov's inequality, we can find

$$P(X \geq 3) \leq \frac{\mathbb{E}(X)}{3} = \frac{8}{9}.$$

   This means that the probability of X being greater than or equal to 3 is less than or equal to $\frac{8}{9}$.

   The precise value can be computed as $P(X \geq 3) = 1 - P(X < 3) = 1 - F(3) = 1 - \int\limits_{-\infty}^{3} f(x)\,dx = 1 - \int\limits_0^3 \frac{x}{8}\,dx = 1 - \frac{x^2}{16}\Big|_0^3 = 1 - \frac{9}{16} = \frac{7}{16}$.

Proof of Markov's Inequality (Optional)

This section is optional and intended for those interested in the justification of Markov's inequality.

We will now prove Markov's inequality for continuous random variables (the same proof works in the discrete case by replacing integrals with sums).

Consider a non-negative continuous random variable X with probability density function $f(x)$. The expected value of X can be written as:

$$\mathbb{E}(X) = \int\limits_0^\infty x f(x)\,dx = \int\limits_0^a x f(x)\,dx + \int\limits_a^\infty x f(x)\,dx.$$

Since $x \geq a$ on the interval $[a, \infty)$, it follows that:

$$\int\limits_a^\infty x f(x)\,dx \geq \int\limits_a^\infty a f(x)\,dx = a \int\limits_a^\infty f(x)\,dx = a\,P(X \geq a).$$

Combining the above, we get:

$$\mathbb{E}(X) \geq a\,P(X \geq a),$$

which implies

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

Chebyshev's inequality

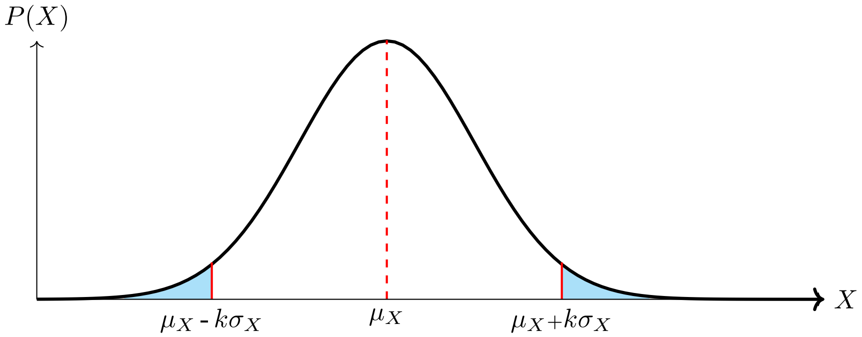
Next we turn our attention to Chebyshev's inequality, a versatile concept with wide-ranging applications. It plays a pivotal role in various fields. In finance, it assists in managing investment risk by estimating the likelihood of significant deviations from expected returns. In manufacturing, it ensures that a substantial majority of products meet quality standards, even when the distribution of characteristics is uncertain. For betting strategies, especially in sports betting, it provides insights into the probability of experiencing specific losses over a given period. In epidemiology and public health, Chebyshev's inequality aids in estimating the proportion of a population at risk of contracting a disease within a defined timeframe, based on the mean and variance of infection rates.

For any random variable X (not necessarily non-negative) with finite variance σ_X^2, and any $k > 0$, Chebyshev's inequality states

$$P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2},$$

where $\mu_X = \mathbb{E}(X)$ is the mean of X.

This inequality provides a bound on the probability of a random variable deviating from its mean by a certain number of standard deviations.



Remark. In order to apply Chebyshev's inequality to estimate the probability that a random variable X deviates from its mean by at least some fixed amount $a > 0$, we observe that if we set $k = \frac{a}{\sigma_X}$, then the inequality reads:

$$P\left(|X - \mu_X| \geq k\sigma_X\right) = P\left(|X - \mu_X| \geq \frac{a}{\sigma_X} \cdot \sigma_X\right) = P(|X - \mu_X| \geq a) \leq \frac{1}{k^2} = \frac{\sigma_X^2}{a^2} = \frac{\text{Var}(X)}{a^2}.$$

In summary, we obtain the useful bound:

$$P(|X - \mu_X| \geq a) \leq \frac{\text{Var}(X)}{a^2}.$$

Example.

1. Let W be a random variable representing the weight of a certain type of fruit. The mean weight of a fruit is 150 grams and the variance is 25 grams. Suppose we would like to estimate the probability of the weight of this type of fruit not falling within the range of $130 - 170$ grams. Chebyshev's inequality can be applied to find

$$P(|W - 150| \geq 20) \leq \frac{25}{20^2} = \frac{1}{4^2} = 0.0625.$$

   This means that the probability of the weight of this type of fruit not falling within the range of $130 - 170$ grams is less than or equal to 6.25%.

2. Suppose X is a random variable such that $\mathbb{E}(X) = 3$ and $\mathbb{E}(X^2) = 13$, we can use Chebyshev's inequality to determine a lower bound for the probability $P(-2 < X < 8)$.

In this case, we have $\mathbb{E}(X) = 3$ and $\mathbb{E}(X^2) = 13$. The variance of X can be calculated using the formula

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 13 - 3^2 = 13 - 9 = 4.$$

We want to find $P(-2 < X < 8)$, which can be rephrased as $P(|X-3| < 5)$. Chebyshev's inequality gives

$$P(|X-3| \geq 5) \leq \frac{4}{5^2} = \frac{4}{25},$$

therefore, we conclude that

$$P(|X-3| < 5) = 1 - P(|X-3| \geq 5) \geq 1 - \frac{4}{25} = \frac{21}{25} \approx 0.84.$$

So, Chebyshev's inequality gives a lower bound of approximately 84% for $P(-2 < X < 8)$.

Markov vs. Chebyshev Inequality: a Comparison

While Chebyshev's inequality applies to deviations from the mean in *both* directions, it can still be used to estimate how frequently a random variable takes on large values. Compared to Markov's inequality, Chebyshev's inequality typically yields tighter bounds when variance information is available.

To illustrate this, let us revisit the earlier example of flipping a biased coin 10 times, where the probability of heads is 20%. Markov's inequality gave an upper bound of $\frac{1}{4}$ on the probability of getting at least 8 heads. We now refine this using Chebyshev's inequality.

Let $\widetilde{X}$ be the random variable defined by $\widetilde{X}(H) = 1$, $\widetilde{X}(T) = 0$. Then:

$$\begin{aligned} \mathbb{E}(\widetilde{X}) &= 0.2 \cdot 1 + 0.8 \cdot 0 = 0.2, \\ \mathrm{Var}(\widetilde{X}) &= \mathbb{E}(\widetilde{X}^2) - (\mathbb{E}(\widetilde{X}))^2 = 0.2 - 0.04 = 0.16. \end{aligned}$$

Let $X = \widetilde{X}_1 + \widetilde{X}_2 + \cdots + \widetilde{X}_{10}$, where the $\widetilde{X}_i$ are independent and identically distributed copies of $\widetilde{X}$. Then:

$$\begin{aligned} \mathbb{E}(X) &= 10 \cdot \mathbb{E}(\widetilde{X}) = 10 \cdot 0.2 = 2, \\ \mathrm{Var}(X) &= 10 \cdot \mathrm{Var}(\widetilde{X}) = 10 \cdot 0.16 = 1.6. \end{aligned}$$

We now apply Chebyshev's inequality to bound:

$$P(|X-2| \geq 6) = P(X \geq 8),$$

since X is non-negative and $X \leq -4$ is impossible.

Chebyshev's inequality gives the following bound:

$$P\left(|X-2| \geq 6\right) \leq \frac{1.6}{6^2} = \frac{1.6}{36} \approx 0.044.$$

Remark. This bound of approximately 4.4% is significantly tighter than the 25% bound obtained via Markov's inequality. This demonstrates the advantage of incorporating variance information via Chebyshev's inequality.

Lecture 13

Adding a Second Variable: Joint and Marginal Distributions

In the upcoming lectures, we will study the behavior of multiple random variables considered together. Many ideas from the single-variable setting will naturally extend to this context. However, new challenges also emerge—particularly due to the possible interactions or dependencies between different variables.

Joint Distributions: Joint PMF

To introduce the idea of a joint distribution, consider the simple experiment of flipping a fair coin three times. Consider two random variables:

- X: the number of heads in the **first two** coin flips,
- Y: the number of heads in the **last two** coin flips.

The sample space consists of the $2^3 = 8$ possible sequences of heads (H) and tails (T). For each outcome, we compute the corresponding values of X and Y:

Flips	X	Y
HHH	2	2
HHT	2	1
HTH	1	1
THH	1	2
HTT	1	0
TTH	0	1
THT	1	1
TTT	0	0

From this, we can compute the distribution of each variable individually (*marginal distribution*). For instance, the marginal distribution of X is:

$$
\begin{aligned}
P(X=0) &= \frac{1}{8} + \frac{1}{8} = \frac{2}{8} = \frac{1}{4}, && \text{(outcomes: TTH, TTT)} \\
P(X=1) &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{4}{8} = \frac{1}{2}, && \text{(outcomes: HTH, THH, HTT, THT)} \\
P(X=2) &= \frac{1}{8} + \frac{1}{8} = \frac{2}{8} = \frac{1}{4}, && \text{(outcomes: HHH, HHT)}
\end{aligned}
$$

But more importantly, we can study the **joint distribution** of X and Y, which describes the probability of each pair (x, y). For example,

$$P(X = 1 \text{ and } Y = 1) = \frac{2}{8} = \frac{1}{4}.$$

The full joint probability mass function is summarized in the table below, where each entry gives the value of $p(x, y) = P(X = x, Y = y)$:

	$Y=0$	$Y=1$	$Y=2$
$X=0$	$1/8$	$1/8$	0
$X=1$	$1/8$	$2/8$	$1/8$
$X=2$	0	$1/8$	$1/8$

This table represents the **joint PMF** of the random variables X and Y. It gives the probabilities of all combinations of values the two variables can simultaneously take.

Two Continuous Distributions: Joint PDF

Just as the behavior of two discrete random variables is described by their joint PMF $p(x, y)$, the behavior of two continuous random variables is captured by their *joint probability density function* $f(x, y)$.

In this context, the probability that X and Y fall within a given rectangle is computed as

$$P(a \le X \le b \text{ and } c \le Y \le d) = \int\limits_a^b \int\limits_c^d f(x, y)\, dy\, dx.$$

More generally, for any subset $S \subseteq \mathbb{R}^2$, we have

$$P((X, Y) \in S) = \iint\limits_S f(x, y)\, dy\, dx.$$

Example. Let (X, Y) have joint PDF $f(x, y) = \begin{cases} x + y & \text{if } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$

Question. What is $P\left(0 < x < \frac{1}{2} \text{ and } 0 < y < \frac{2}{3}\right)$?

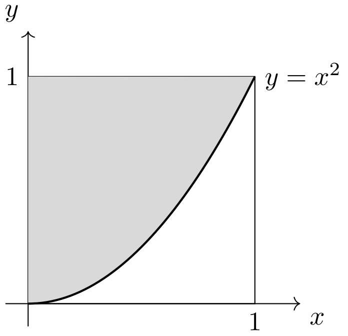
This is the integral of the joint density over the specified rectangle:

$$\int_0^{1/2}\int_0^{2/3} (x+y)\,dy\,dx = \int_0^{1/2} \left(xy + \frac{y^2}{2}\right)\Bigg|_{y=0}^{2/3} dx = \int_0^{1/2} \left(\frac{2}{9} + \frac{2}{3}x\right) dx = \left(\frac{2}{9}x + \frac{1}{3}x^2\right)\Bigg|_0^{0.5} = \frac{7}{36}.$$

Remark. In continuous settings, the distinction between strict ($<$) and non-strict ($\leq$) inequalities does not affect the probability values.

Question. What is $P(Y \geq X^2)$?

We now compute the probability that the point (X, Y) lies in the region above the parabola $y = x^2$ and within the unit square:



The probability under consideration can be calculated as

$$P(Y \geq X^2) = \int_0^1\int_{x^2}^1 (x+y)\,dy\,dx = \int_0^1 \left(xy + \frac{y^2}{2}\right)\Bigg|_{y=x^2}^{1} dx = \int_0^1 \left(x + \frac{1}{2} - \left(x^3 + \frac{x^4}{2}\right)\right) dx =$$

$$\left(\frac{x^2}{2} + \frac{x}{2} - \frac{x^4}{4} - \frac{x^5}{10}\right)\Bigg|_0^1 = 0.5 + 0.5 - 0.25 - 0.1 = 0.65.$$

Joint CDF

Recall that for a single continuous random variable X, the cumulative distribution function is defined as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt.$$

When dealing with two continuous random variables (X, Y), we define the **joint CDF** analogously. It is now a function of two variables:

$$F(x, y) = P(X \leq x \text{ and } Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(s, t)\,dt\,ds.$$

Remark. As in the single-variable case, the integral may not actually start at $-\infty$ depending on where the PDF is nonzero. Also, remember that when you do the integration, you are treating x and y as constants!

Example. Let (X, Y) have the same joint PDF from the previous example:

$$f(x,y) = \begin{cases} x + y & \text{if } 0 \leq x \leq 1,\ 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

We compute the joint CDF $F(x, y) = P(X \leq x \text{ and } Y \leq y)$ for $0 \leq x \leq 1$ and $0 \leq y \leq 1$:

$$F(x,y) = \int_0^x \int_0^y (s+t)\, dt\, ds.$$

First, compute the inner integral:

$$\int_0^y (s+t)\, dt = \left(st + \frac{t^2}{2} \right)\Bigg|_{t=0}^{y} = sy + \frac{y^2}{2}.$$

Now integrate with respect to s:

$$\int_0^x \left(sy + \frac{y^2}{2} \right) ds = \left(\frac{s^2 y}{2} + \frac{y^2 s}{2} \right)\Bigg|_0^x = \frac{x^2 y}{2} + \frac{xy^2}{2}.$$

So the joint CDF is

$$F(x,y) = \frac{x^2 y}{2} + \frac{xy^2}{2}, \quad \text{for } 0 \leq x \leq 1,\ 0 \leq y \leq 1.$$

Verification. We can now confirm the probability of a rectangle, such as:

$$P\left(0 < x < \frac{1}{2} \text{ and } 0 < y < \frac{2}{3}\right) = F\left(\frac{1}{2}, \frac{2}{3}\right).$$

$$F\left(\frac{1}{2}, \frac{2}{3}\right) = \frac{(1/2)^2 \cdot (2/3)}{2} + \frac{(1/2) \cdot (2/3)^2}{2} = \frac{1}{12} + \frac{1}{9} = \frac{7}{36},$$

which matches our earlier result.

Marginal Distribution

Occasionally, we are given the joint PMF or PDF of two random variables but are interested in understanding the individual distributions of X and Y. These individual distributions are referred to as *marginal* PMFs or PDFs, and this lecture is devoted to exploring how they can be extracted from the joint distribution.

Discrete Random Variables

To compute the marginal PMF of a discrete random variable, we sum the joint probabilities over all possible values of the other variable.

Example. Consider a bag containing six marbles labeled as:

$$\{(\text{red}, 1), (\text{blue}, 1), (\text{green}, 1), (\text{red}, 2), (\text{blue}, 2), (\text{green}, 2)\}.$$

Let X denote the color and Y the number written on a randomly selected marble. The joint PMF $p(x, y)$ is given in the table below:

x/y	1	2
red	$\frac{1}{6}$	0
blue	$\frac{2}{6}$	$\frac{1}{6}$
green	$\frac{1}{6}$	$\frac{1}{6}$

To compute the **marginal** PMFs of X and Y, we sum over the corresponding rows or columns of the joint PMF:

$$P(X=\text{red})=\sum_{y\in Y}p(\text{red},y)=\frac{1}{6}+0=\frac{1}{6};$$
$$P(X=\text{blue})=\sum_{y\in Y}p(\text{blue},y)=\frac{2}{6}+\frac{1}{6}=\frac{3}{6};$$
$$P(X=\text{green})=\sum_{y\in Y}p(\text{green},y)=\frac{1}{6}+\frac{1}{6}=\frac{2}{6};$$
$$P(Y=1)=\sum_{x\in X}p(x,1)=\frac{1}{6}+\frac{2}{6}+\frac{1}{6}=\frac{4}{6};$$
$$P(Y=2)=\sum_{x\in X}p(x,2)=0+\frac{1}{6}+\frac{1}{6}=\frac{2}{6}.$$

Continuous Random Variables

Similarly, if $f(x,y)$ is the joint probability density function of continuous variables X and Y, the marginal PDFs of X are given by:

$$f_X(x)=\int\limits_{-\infty}^{\infty}f(x,y)\,dy$$
$$f_Y(y)=\int\limits_{-\infty}^{\infty}f(x,y)\,dx.$$

To remember this intuitively, consider the following ”Probability Party” slogan:

To find the marginal distribution, sum/integrate out the variable you don’t want to keep.

Example. Let X and Y have the joint PDF $f(x,y)=15x^2y$, $0<x<y<1$, and zero elsewhere.

First let’s check that $f(x,y)$ is indeed a valid PDF:

- $f(x,y)\geq 0$ for any $(x,y)\in\mathbb{R}^2$
- $\iint\limits_{\mathbb{R}^2}f(x,y)\,dy\,dx=\int\limits_0^1\int\limits_x^1 15x^2\cdot y\,dy\,dx=\frac{15}{2}\int\limits_0^1 x^2y^2\Big|_{y=x}^1\,dx=\frac{15}{2}\int\limits_0^1 x^2(1-x^2)\,dx=\frac{15}{2}\int\limits_0^1(x^2-x^4)\,dx=\frac{15}{2}\left(\frac{x^3}{3}-\frac{x^5}{5}\right)\Big|_0^1=\frac{15}{2}\left(\frac{1}{3}-\frac{1}{5}\right)=\frac{15}{2}\cdot\frac{2}{15}=1$ ✓

Next we compute the marginal PDF of X:

$$f_X(x)=\int\limits_{-\infty}^{\infty}f(x,y)\,dy=\int\limits_x^1 15x^2y\,dy=15x^2\int\limits_x^1 y\,dy=15x^2\frac{y^2}{2}\Big|_x^1=7.5x^2(1-x^2),$$

where $0<x<1$.

Now, let’s find the marginal PDF of Y:

$$f_Y(y)=\int\limits_{-\infty}^{\infty}f(x,y)\,dx=\int\limits_0^y 15x^2y\,dx=15y\int\limits_0^y x^2\,dx=15y\cdot\frac{x^3}{3}\Big|_0^y=5y^4,$$

where $0<y<1$.

So, the marginal PDFs are

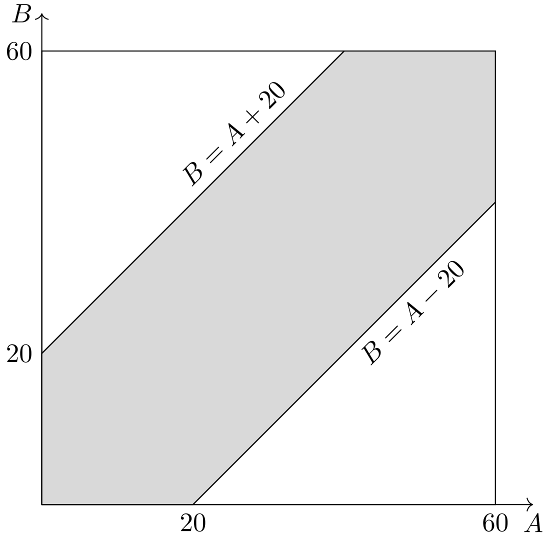
$$f_X(x) = \begin{cases} \frac{15}{2}x^2(1-x^2), & 0 < x < 1, \\ 0, & \text{otherwise} \end{cases} \quad \text{and } f_Y(y) = \begin{cases} 5y^4, & 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Now, let's deviate for a moment and see how what we have learned about the probability theory of two random variables can be applied to a real-life situation.

Example. Alice and Bob plan to meet at a train station between 9 and 10 am, but they haven't agreed on a specific time. Each of them arrives at some time between 9 and 10 am, with all times being equally likely. After waiting for 20 minutes, they give up and head home. What is the probability that they actually meet?

To formalize this scenario, let's consider the random variables (A, B), representing the arrival times of Alice and Bob, respectively. These variables are uniformly distributed on the square $0 \leq A \leq 60$ and $0 \leq B \leq 60$, where the unit of time is in minutes. Alice and Bob will meet if and only if the absolute difference between their arrival times satisfies $|A - B| \leq 20$.

In other words, we need to find the probability that $|A - B| \leq 20$. The visual representation of this condition is illustrated below, where our objective is to determine the probability that the pair (A, B) falls within the shaded region:



The total area of the square is $60 \times 60 = 3600$. To find the area of the shaded region, we subtract the areas of the two right triangles from the total area.

Each triangle has a base and height of 40, so the cumulative area is

$$2 \cdot \frac{1}{2} \cdot 40 \cdot 40 = 1600.$$

Therefore, the area of the shaded region is $3600 - 1600 = 2000$. The probability corresponding to the shaded region is the ratio of its area to the total area $\frac{2000}{3600} = \frac{5}{9}$.

Applications of Marginal Distributions

Marginal distributions are not just a mathematical concept — they have widespread applications in a variety of real-world domains.

- **Artificial Intelligence & Machine Learning.** Marginal distributions are central to probabilistic models such as Bayesian networks, where the probability of an outcome is computed by summing/integrating over hidden or latent variables. For example, in natural language processing, one might compute the marginal probability of a word occurring in a sentence, regardless of the underlying grammatical structure.

- **Economics.** In econometrics, marginal distributions help in understanding the behavior of individual economic variables, such as income or consumption, when data is collected jointly. For instance, given joint data on education level and income, the marginal distribution of income can help in tax policy design or welfare analysis.
- **Finance.** Risk analysts use marginal distributions to evaluate the distribution of returns for individual assets when analyzing portfolios. This helps in stress testing and risk management.
- **Operations Research.** In logistics and supply chain optimization, marginal probabilities help model demand at individual locations based on a joint distribution over regions or products.

In summary, marginal distributions allow us to focus on the behavior of a single variable while appropriately accounting for dependencies or interactions with other variables.

Lecture 14
Two Random Variables: Expectations and Transformations

Let X and Y be two discrete or continuous random variables with joint PMF $p(x, y)$ (in the discrete case) or joint PDF $f(x, y)$ (in the continuous case). Suppose $Z = g(X, Y)$ for some real-valued function $g : \mathbb{R}^2 \to \mathbb{R}$. Then Z is itself a random variable, and its expected value can be computed as follows.

Definition. If (X, Y) are continuous random variables, the expected value of Z is given by

$$\mathbb{E}(Z) = \iint\limits_{\mathbb{R}^2} g(x, y) f(x, y)\, dy\, dx,$$

provided the integral of the absolute value converges:

$$\iint\limits_{\mathbb{R}^2} |g(x, y)| f(x, y)\, dy\, dx < \infty.$$

If (X, Y) are discrete random variables, the expected value of Z is given by

$$\mathbb{E}(Z) = \sum_x \sum_y g(x, y) p(x, y),$$

assuming the double sum of absolute values converges:

$$\sum_x \sum_y |g(x, y)| p(x, y) < \infty.$$

Remark. When dealing with infinite series (such as expectations expressed as infinite sums), it is crucial to distinguish between *absolute* and *conditional* convergence. For instance, a remarkable result due to Riemann, known as the *Riemann Rearrangement Theorem*, states that if a series converges conditionally—that is, it converges but not absolutely—then its terms can be rearranged to make the series converge to any real number, or even to diverge.

This has serious implications for the computation of expectations: if the sum defining the expected value is only conditionally convergent, then the expected value is not well-defined, since its value may depend on the order in which we sum the terms. In contrast, absolute convergence guarantees that any rearrangement yields the same result, ensuring the expectation is meaningful.

Optional: Example of Rearranging a Conditionally Convergent Series

To illustrate the pathological nature of conditionally convergent series, we include the following classical example.

Example. Suppose we wish to arrange the signs in the series

$$\sum_{n=1}^{\infty} \frac{a_n}{n},$$

where each $a_n \in \{-1, 1\}$, so that the resulting series converges to 2. One way to do this is to iteratively add or subtract terms in order to ”steer” the partial sums toward 2.

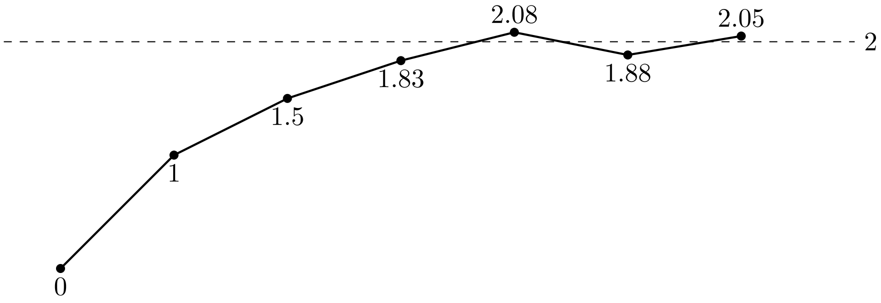
Start with $a_1 = 1 \Rightarrow S_1 = 1$. Since $S_1 < 2$, continue increasing:

$$S_2 = 1 + \frac{1}{2} = 1.5, \quad S_3 = 1.5 + \frac{1}{3} \approx 1.833, \quad S_4 = 1.833 + \frac{1}{4} \approx 2.083.$$

Now that we have overshot 2, we begin subtracting:

$$S_5 = 2.083 - \frac{1}{5} = 1.883, \quad S_6 = 1.883 + \frac{1}{6} \approx 2.05.$$

Continuing this way—alternating signs when we cross the target value—we can construct a sequence $\{a_n\}$ that makes the series converge to exactly 2.



This construction illustrates how a conditionally convergent series can be manipulated to achieve a desired sum, underscoring the importance of absolute convergence in probability and analysis.

Linearity of Expectation

- If $W = g(X, Y)$ and $Z = h(X, Y)$ are two real-valued functions of X and Y, then

$$\mathbb{E}(W + Z) = \mathbb{E}(W) + \mathbb{E}(Z).$$

- For any constant $c \in \mathbb{R}$, we have

$$\mathbb{E}(cW) = c\mathbb{E}(W).$$

Example. Let X and Y be discrete random variables taking values in $\{1, 2\}$, and suppose their joint PMF is given by

$$p(x, y) = \frac{x + y}{12}.$$

Compute $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathbb{E}(X^2)$, $\mathbb{E}(XY)$, and determine whether $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. Then compute $\mathbb{E}(2X^2 - 6XY + 7Y)$.

Solution. To compute the expected values, we use the definitions:

$$\mathbb{E}(X) = \sum_x \sum_y x \cdot p(x,y) = \sum_{x=1}^{2} \sum_{y=1}^{2} x \cdot p(x,y) = 1 \cdot \left(\frac{1+1}{12} + \frac{1+2}{12} \right) + 2 \cdot \left(\frac{2+1}{12} + \frac{2+2}{12} \right) = \frac{19}{12};$$

$$\mathbb{E}(Y) = \sum_x \sum_y y \cdot p(x,y) = \sum_{x=1}^{2} \sum_{y=1}^{2} y \cdot p(x,y) = 1 \cdot \left(\frac{1+1}{12} + \frac{1+2}{12} \right) + 2 \cdot \left(\frac{2+1}{12} + \frac{2+2}{12} \right) = \frac{19}{12};$$

$$\mathbb{E}(X^2) = \sum_x \sum_y x^2 \cdot p(x,y) = \sum_{x=1}^{2} \sum_{y=1}^{2} x^2 \cdot p(x,y) = 1^2 \cdot \left(\frac{1+1}{12} + \frac{1+2}{12} \right) + 2^2 \cdot \left(\frac{2+1}{12} + \frac{2+2}{12} \right) = \frac{33}{12};$$

$$\mathbb{E}(XY) = \sum_x \sum_y xy \cdot p(x,y) = 1 \cdot 1 \cdot \frac{1+1}{12} + 2 \cdot 1 \cdot \frac{1+2}{12} + 1 \cdot 2 \cdot \frac{2+1}{12} + 2 \cdot 2 \cdot \frac{2+2}{12} = \frac{30}{12}.$$

Next we check that $\mathbb{E}(X)\mathbb{E}(Y) \neq \mathbb{E}(XY)$.

Finally, we use linearity of expectation to compute $\mathbb{E}(2X^2 - 6XY + 7Y) = 2\mathbb{E}(X^2) - 6\mathbb{E}(XY) + 7\mathbb{E}(Y) = 2 \cdot \frac{33}{12} - 6 \cdot \frac{30}{12} + 7 \cdot \frac{19}{12} = \frac{19}{12}$.

Example. Consider random variables X and Y with a joint probability density function given by:

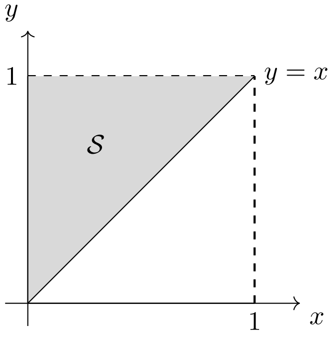
$$f(x,y) = \begin{cases} 8xy & \text{if } 0 \leq x < y \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

We aim to calculate the expected value of $g(x,y) = x^2 - y$. This involves the double integral:

$$\mathbb{E}(g(x,y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y)\, dy\, dx = \iint_{\mathcal{S}} 8(x^2 - y)xy\, dy\, dx,$$

where $\mathcal{S} := \{(x,y) \in \mathbb{R}^2 \mid 0 < x < y < 1\}$ represents the region of integration.

Visualizing this region in the coordinate plane, we have a triangular area above the line $y = x$, bounded by $0 < x < 1$ and $0 < y < 1$:



Now, we express the integral as $\int_0^1 \int_x^1 8(x^2 - y)xy\, dy\, dx$. Evaluating the inner integral yields

$$\int_x^1 8(x^2 - y)xy\, dy = \int_x^1 (8x^3 y - 8xy^2)\, dy = \left(4x^3 y^2 - \frac{8}{3} x y^3 \right) \Bigg|_{y=x}^{y=1} = 4x^3 - \frac{8}{3}x - 4x^5 + \frac{8}{3}x^4.$$

Substituting this into the outer integral, we obtain

$$\int_0^1 \left(4x^3 - \frac{8}{3}x - 4x^5 + \frac{8}{3}x^4 \right) dx = \left(x^4 - \frac{4}{3}x^2 - \frac{4}{6}x^6 + \frac{8}{15}x^5 \right) \Bigg|_0^1 = 1 - \frac{4}{3} - \frac{4}{6} + \frac{8}{15} = -\frac{7}{15}.$$

Therefore, the expected value of $g(x,y)$ is $-\frac{7}{15}$.

Transforming Pairs of Random Variables

Consider two random variables (X_1, X_2) with a known joint probability density function. However, our primary interest lies in two other variables, (Y_1, Y_2), expressed as functions of X_1 and X_2:

$$Y_1 = u_1(X_1, X_2)$$

$$Y_2 = u_2(X_1, X_2).$$

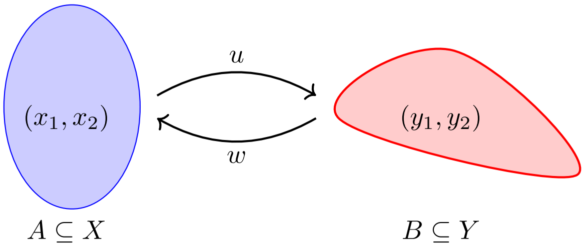
Our goal is to determine the joint PDF of the Y variables, denoted as $g(y_1, y_2)$. By definition, this function should satisfy the equation

$$P((Y_1, Y_2) \in B) = \iint\limits_B g(y_1, y_2)\, dy_2\, dy_1.$$

This will be accomplished with the help of the equality

$$P((Y_1, Y_2) \in B) = P((X_1, X_2) \in w(B) = A).$$

Assuming the transformation $u = (u_1, u_2) : (x_1, x_2) \to (y_1, y_2)$ is one-to-one, we can invert it. This allows us to determine functions w_1 and w_2 such that $X_1 = w_1(Y_1, Y_2)$ and $X_2 = w_2(Y_1, Y_2)$:



Joint PDF of Transformed Random Variables, a.k.a. Cobbling the Jacobian 🙂

Before we study multivariable change-of-variable techniques using the Jacobian, let's briefly recall the idea of single-variable u-substitution. Suppose we have an integral of the form

$$\int f(x)\, dx,$$

and we make the substitution $u = g(x)$, where g is a differentiable and invertible function. Then the differential element transforms as

$$dx = \frac{\partial x}{\partial u} \cdot du = \frac{1}{g'(x)} \cdot du.$$

Thus, the integral becomes

$$\int f(x)\, dx = \int f(x(u)) \cdot \left| \frac{\partial x}{\partial u} \right| du = \int f(g^{-1}(u)) \cdot \left| \frac{1}{g'(x)} \right| du.$$

This idea generalizes naturally to the multivariable case, where the Jacobian matrix contains partial derivatives and its determinant scales area (in two dimensions) or volume (in higher dimensions) under coordinate transformations.

Example. In order to evaluate the integral $\int x(5x^2 - 24)^{19}\, dx$,
we will use the substitution

$$u = 5x^2 - 24 \text{ with } du = \frac{\partial u}{\partial x} \cdot dx = 10x\, dx \Leftrightarrow dx = \frac{1}{10x} \cdot du.$$

Substituting into the original integral allows to find:

$$\int x(5x^2-24)^{19}\,dx = \int u^{19}\cdot\frac{1}{10}\,du = \frac{1}{10}\int u^{19}\,du.$$

Now integrate:

$$\frac{1}{10}\cdot\frac{u^{20}}{20}+C = \frac{1}{200}u^{20}+C.$$

Finally, we substitute back $u = 5x^2 - 24$ to get

$$\int x(5x^2-24)^{19}\,dx = \frac{1}{200}(5x^2-24)^{20}+C.$$

Remark. In this case, the substitution $u = 5x^2 - 24$ introduces a scaling factor of $\frac{\partial x}{\partial u} = \frac{1}{\frac{du}{dx}} = \frac{1}{10x}$, which plays the role of the Jacobian of the *inverse transformation.* The Jacobi matrix is a 1×1 matrix whose sole entry is $\frac{\partial x}{\partial u}$, and its determinant is equal to that entry.

This idea generalizes to the multivariable case, where the *Jacobian* corrects for the distortion of area or volume element under a coordinate transformation. The Jacobian of a transformation $u : (x_1, x_2) \to (y_1, y_2) = (u_1(x_1, x_2), u_2(x_1, x_2))$ is the determinant of the 2×2 matrix of partial derivatives denoted by $J(u)$. It represents the scaling factor by which the transformation affects the area element when changing from (x_1, x_2) coordinates to (y_1, y_2) coordinates:

$$J(u) = \begin{pmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} \\ \frac{\partial u_2}{\partial x_1} & \frac{\partial u_2}{\partial x_2} \end{pmatrix}.$$

When computing double integrals in the transformed space, the Jacobian appears in the differential element as $|J(u)|\,dy_2\,dy_1$. It ensures that area is conserved between the two spaces.

Let's consider an example of a transformation and compute its Jacobian.

Example. Let u be the transformation from polar coordinates (r, φ) to cartesian coordinates $(x, y) = (r\cos(\varphi), r\sin(\varphi))$. The Jacobian determinant for the given transformation from polar coordinates to Cartesian coordinates is as follows.

First we compute the partial derivatives of new coordinates with respect to old:

$$\frac{\partial x}{\partial r} = \cos(\varphi), \quad \frac{\partial x}{\partial \varphi} = -r\sin(\varphi)$$
$$\frac{\partial y}{\partial r} = \sin(\varphi), \quad \frac{\partial y}{\partial \varphi} = r\cos(\varphi)$$

and record then in the matrix of partial derivatives:

$$J = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \varphi} \end{pmatrix} = \begin{pmatrix} \cos(\varphi) & -r\sin(\varphi) \\ \sin(\varphi) & r\cos(\varphi) \end{pmatrix}$$

Finally, we compute the Jacobian as

$$\det(J) = \cos(\varphi)\cdot r\cos(\varphi) - (-r\sin(\varphi))\cdot\sin(\varphi)) = r\cos^2(\varphi) + r\sin^2(\varphi) = r.$$

Hence, when changing variables in a double integral from Cartesian to polar coordinates, we multiply by the Jacobian determinant r.

One important use of the polar coordinate transformation is in proving that the probability density function for the standard normal distribution integrates to 1:

$$\int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx = 1.$$

Let us verify that this is indeed the case.

Let us denote this integral by A, and show that $A^2 = 1$. We write A^2 as

$$A^2 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy = \frac{1}{2\pi} \iint_{\mathbb{R}^2} e^{-x^2/2} e^{-y^2/2}\, dy\, dx = \frac{1}{2\pi} \iint_{\mathbb{R}^2} e^{(-x^2-y^2)/2}\, dy\, dx.$$

We switch to polar coordinates using $x^2 + y^2 = r^2$, and recall the Jacobian from earlier:

$$A^2 = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2}\, r dr\, d\varphi = \frac{1}{2\pi} \int_0^{2\pi} \left(-e^{-r^2/2} \Big|_0^{\infty} \right) d\varphi = \frac{1}{2\pi} \int_0^{2\pi} d\varphi = \frac{1}{2\pi} \cdot 2\pi = 1 \checkmark$$

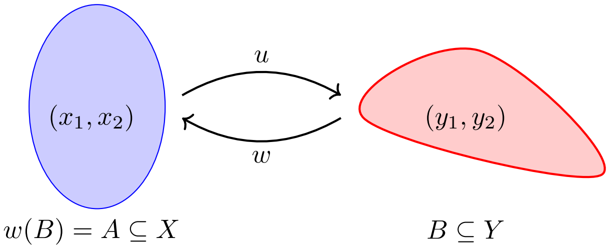
Remark. When changing variables in a double integral—such as converting from Cartesian coordinates (x, y) to polar coordinates (r, φ)—we compute the *Jacobian of the inverse transformation*, that is, the determinant of the matrix of partial derivatives of the **original variables** (x, y) with respect to the **new variables** (r, φ).

We will now use the equality of probabilities

$$P\big((Y_1, Y_2) \in B\big) = P\big((X_1, X_2) \in w(B)\big),$$

to express the joint density function of (Y_1, Y_2) in terms of the density of (X_1, X_2). Let $u = (u_1, u_2) : (x_1, x_2) \mapsto (y_1, y_2)$ is a one-to-one, continuously differentiable transformation with inverse $w = (w_1, w_2)$ such that

$$x_1 = w_1(y_1, y_2), \quad x_2 = w_2(y_1, y_2).$$



By substituting the right-hand side using the density function f of (X_1, X_2), and applying the change-of-variables formula, we obtain:

$$\iint_B g(y_1, y_2)\, dy_1 dy_2 = \iint_B f(w_1(y_1, y_2), w_2(y_1, y_2)) \cdot |\det J_w(y_1, y_2)|\, dy_1 dy_2.$$

Theorem. Let (X_1, X_2) be continuous random variables with joint probability density function $f(x_1, x_2)$. Suppose $(Y_1, Y_2) = u(X_1, X_2)$ are new variables defined by a one-to-one, continuously differentiable transformation $u : \mathbb{R}^2 \to \mathbb{R}^2$, with inverse $w = u^{-1}$. Let $J_w(y_1, y_2)$ be the Jacobian matrix of partial derivatives of w evaluated at (y_1, y_2). Then the joint probability density function g of (Y_1, Y_2) is given by

$$g(y_1, y_2) = f(w_1(y_1, y_2), w_2(y_1, y_2)) \cdot |\det J_w(y_1, y_2)|\, .$$

Remark. The formula above expresses how a change of variables affects a probability density. To interpret it intuitively: evaluate the original density f at the pre-image $(x_1, x_2) = w(y_1, y_2)$, then multiply by the absolute value of the Jacobian determinant of the inverse transformation. In short:

Density at new coordinates = original density (at old coords) × scaling factor (Jacobian).

Example. Let (X_1, X_2) be a pair of continuous random variables with joint probability density function

$$f(x_1, x_2) = \begin{cases} 2x_1^3 x_2, & 0 < x_1 < 1,\ 0 < x_2 < 2, \\ 0, & \text{otherwise.} \end{cases}$$

Consider the change of variables $u : \mathbb{R}^2 \to \mathbb{R}^2$ defined by the linear transformation:

$$\begin{cases} y_1 = 8x_1 - 2x_2, \\ y_2 = -2x_1 + x_2. \end{cases}$$

Our goal is to find the joint PDF of the pair of random variables (Y_1, Y_2).

Step 1. Find the inverse transformation.

This can be done either by solving the system algebraically or by computing the inverse of the coefficient matrix

$$A = \begin{pmatrix} 8 & -2 \\ -2 & 1 \end{pmatrix}, \quad A^{-1} = \frac{1}{4}\begin{pmatrix} 1 & 2 \\ 2 & 8 \end{pmatrix} = \begin{pmatrix} 0.25 & 0.5 \\ 0.5 & 2 \end{pmatrix}.$$

Hence, the inverse transformation $w = u^{-1}$ is given by:

$$\begin{cases} x_1 = w_1(y_1, y_2) = 0.25y_1 + 0.5y_2, \\ x_2 = w_2(y_1, y_2) = 0.5y_1 + 2y_2. \end{cases}$$

Step 2. Compute the Jacobian determinant of the inverse transformation.

The Jacobian matrix is

$$J(w) = \begin{pmatrix} \frac{\partial w_1}{\partial y_1} & \frac{\partial w_1}{\partial y_2} \\ \frac{\partial w_2}{\partial y_1} & \frac{\partial w_2}{\partial y_2} \end{pmatrix} = \begin{pmatrix} 0.25 & 0.5 \\ 0.5 & 2 \end{pmatrix}, \quad \text{so} \quad |\det(J(w))| = 0.25 \cdot 2 - 0.5 \cdot 0.5 = 0.25.$$

Step 3. Compute the joint PDF of (Y_1, Y_2).

By the change of variables formula:

$$g(y_1, y_2) = |\det(J(w))| \cdot f(w_1(y_1, y_2), w_2(y_1, y_2)) = 0.25 \cdot f(0.25y_1 + 0.5y_2,\ 0.5y_1 + 2y_2).$$

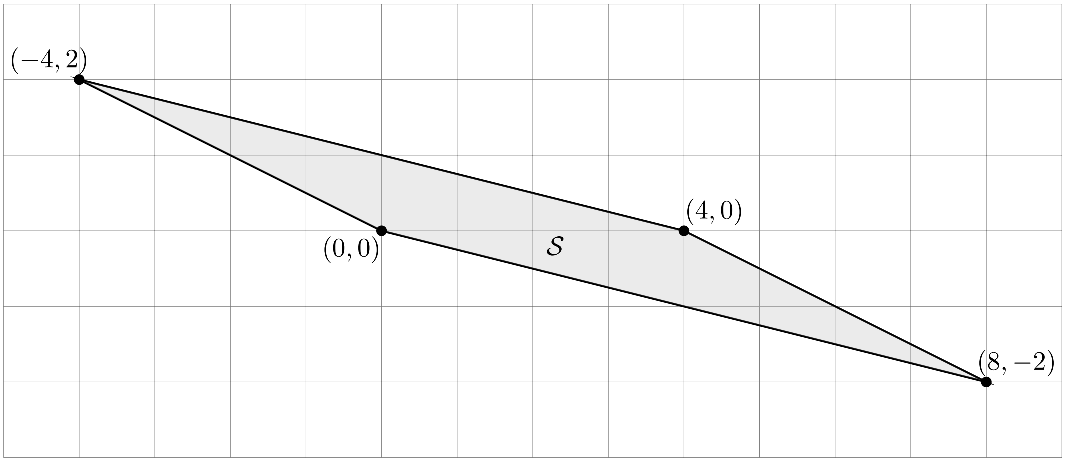
Since $f(x_1, x_2) = 2x_1^3 x_2$ inside the rectangle $0 < x_1 < 1,\ 0 < x_2 < 2$, we substitute:

$$g(y_1, y_2) = \begin{cases} 0.25 \cdot 2(0.25y_1 + 0.5y_2)^3(0.5y_1 + 2y_2), & (y_1, y_2) \in \mathcal{S}, \\ 0, & \text{otherwise,} \end{cases}$$

which simplifies to:

$$g(y_1, y_2) = \begin{cases} 0.5(0.25y_1 + 0.5y_2)^3(0.5y_1 + 2y_2), & (y_1, y_2) \in \mathcal{S}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathcal{S}$ is the image of the rectangle $0 < x_1 < 1,\ 0 < x_2 < 2$ under the transformation u. This region $\mathcal{S}$ is a parallelogram with vertices at $(0, 0), (4, 0), (8, -2), (-4, 2)$:



Remark. Since u is a linear transformation, the image of the rectangle $0 < x_1 < 1,\ 0 < x_2 < 2$ under u is a parallelogram $\mathcal{S}$. The area of this parallelogram is equal to the absolute value of the determinant of the Jacobian matrix of the forward transformation u, that is:

$$\left|\det\begin{pmatrix} 8 & -2 \\ -2 & 1 \end{pmatrix}\right| = |8 \cdot 1 - (-2) \cdot (-2)| = |8 - 4| = 4.$$

Lecture 15

PDF of Sum: Not That Convoluted

Often in probability and statistics, we encounter situations where the variable of interest can be decomposed into the sum of several effects:

$$Y = (\text{contribution to } Y \text{ from effect 1}) + (\text{contribution to } Y \text{ from effect 2}) + \dots$$

For instance, the total revenue for a business day could be the accumulation of revenues from various sources, such as sales, services, and additional income streams. Similarly, in a different context, the overall energy consumption in a household throughout a day might result from the sum of energy usage from different appliances and devices, including lighting, heating, and electronic equipment.

For simplicity, let's assume that there are only two effects at play, and they both have continuous distributions. Furthermore, we assume that we know the distribution of each effect and how these distributions interact with each other. Mathematically, this assumption means that we know the joint probability density function $f(x_1, x_2)$ of two random variables, X_1 and X_2. Our goal is to determine the distribution of $X_1 + X_2$.

PDF of Sum of Two Random Variables

To compute the probability density function of the sum $Y_1 = X_1 + X_2$, we would like to use the variable transformation technique introduced in the previous lecture. However, this technique requires a transformation that outputs two variables. To accommodate this, we introduce an auxiliary (or "dummy") variable $Y_2 := X_2$, and consider the transformation

$$\begin{cases} Y_1 = X_1 + X_2 \\ Y_2 = X_2. \end{cases}$$

Our goal is to determine the PDF of Y_1 by first finding the joint PDF of (Y_1, Y_2), and then marginalizing (integrating out) the superfluous variable Y_2.

To carry this out, we express the original variables (X_1, X_2) in terms of the new variables (Y_1, Y_2). We express (X_1, X_2) in terms of (Y_1, Y_2) using the above equations:

$$\begin{cases} X_1 = Y_1 - Y_2 \\ X_2 = Y_2. \end{cases}$$

The Jacobian determinant of this transformation is

$$\det\begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix} = \det\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = 1.$$

Hence, the joint PDF of (Y_1, Y_2) is given by

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(y_1 - y_2, y_2).$$

To obtain the PDF of Y_1, we integrate out y_2:

$$f_{Y_1}(y_1) = \int\limits_{-\infty}^{\infty} f_{X_1,X_2}(y_1 - y_2, y_2)\, dy_2.$$

Now consider a special case where the random variables X_1 and X_2 are independent. In that case, their joint PDF factorizes:

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2).$$

Substituting into the integral above, we get

$$f_{Y_1}(y_1) = \int\limits_{-\infty}^{\infty} f_{X_1}(y_1 - y_2) \cdot f_{X_2}(y_2)\, dy_2.$$

To compute the probability density function of a sum of two continuous random variables X and Y, we use the formula:

$$f_Z(z) = \int\limits_{-\infty}^{\infty} f_X(z - y) \cdot f_Y(y)\, dy.$$

Remark. In this independent case, the formula for the PDF of the sum $Z = X + Y$ becomes the *convolution* of the individual PDFs:

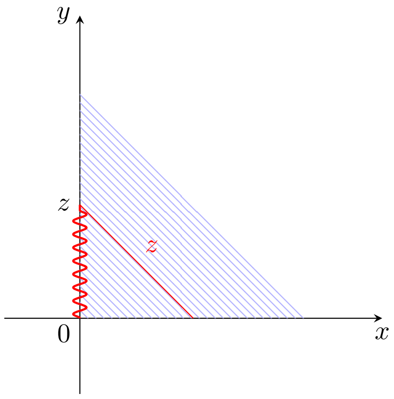
$$(f_X \star f_Y)(z) := \int\limits_{-\infty}^{\infty} f_X(z - y) \cdot f_Y(y)\, dy.$$

The convolution operation, denoted by $\star$, is a fundamental tool in probability and analysis. It reflects how distributions "add together" when random variables are summed.

Example. 1. Consider two random variables X and Y with the joint probability density function

$$f(x, y) = \begin{cases} e^{-(x+y)} & \text{if } x \geq 0 \text{ and } y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

We are interested in computing the PDF of the sum $Z = X + Y$. Note that the joint PDF is supported on the first quadrant, where both x and y are nonnegative. For a fixed value of $z = x + y$, the set of points (x, y) such that $x + y = z$ and $x, y \geq 0$ corresponds to the line segment connecting $(0, z)$ to $(z, 0)$:



To find the PDF of $Z = X + Y$, we integrate out one of the variables over the region where $x + y = z$ and $x, y \geq 0$. For a fixed $z \geq 0$, the variable y ranges from 0 to z, and $x = z - y$. Therefore,

$$f_Z(z) = \int_0^z f(z - y, y)\, dy = \int_0^z e^{-(z-y+y)}\, dy = \int_0^z e^{-z}\, dy = e^{-z} \int_0^z dy = ze^{-z}.$$

Hence, the PDF of $Z = X + Y$ is

$$f_Z(z) = \begin{cases} ze^{-z} & \text{if } z \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

2. Let X be uniformly distributed on $[0, 1]$ with PDF $f_X(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise,} \end{cases}$ and Y be exponentially distributed with parameter λ: $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$. We assume X and Y are independent, so their joint density is the product:

$$f(x, y) := f_X(x) \cdot f_Y(y).$$

To compute the distribution of $Z = X + Y$, we use the convolution formula:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z - y) \cdot f_Y(y)\, dy.$$

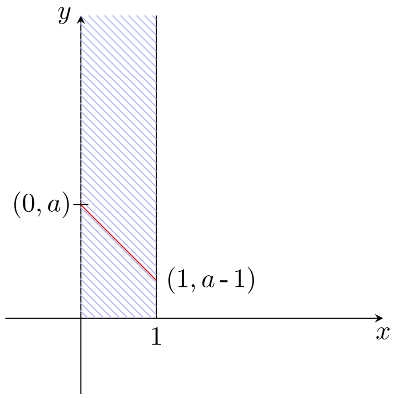
Since $f_X(z - y)$ is nonzero only when $0 \leq z - y \leq 1$, i.e., when $z - 1 \leq y \leq z$, we may rewrite the integral as

$$f_{X+Y}(z) = \int_{\max(0,z-1)}^{z} \lambda e^{-\lambda y}\, dy,$$

valid for $z > 0$. This gives:

$$f_{X+Y}(z) = \begin{cases} e^{-\lambda(z-1)} - e^{-\lambda z} & \text{if } z > 1, \\ 1 - e^{-\lambda z} & \text{if } 0 \leq z \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

The figure below illustrates the support of the joint distribution and the level curves $Z = a$ of the random variable $Z = X + Y$, shown in blue. The diagonal line segment corresponds to points (x, y) such that $x + y = a$, restricted to the strip where $0 \leq x \leq 1$ and $y \geq 0$:



Sum of Fish in the Pond of Probabilities

Before we move on, it is worthwhile to explore an instance of convolution featuring two discrete random variables. The *Poisson distribution* is used to model the number of times an event occurs in a fixed interval of time, under the assumption that events occur independently and at a constant average rate. This rate is denoted by $\lambda > 0$.

The sample space consists of non-negative integers $\{0, 1, 2, \ldots\}$, and the probability mass function of a Poisson random variable X is given by:

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

Remark. The word "Poisson" in *Poisson distribution* comes from the French mathematician Siméon Denis Poisson, but it also happens to mean "fish" in French. This coincidence has inspired the title of the subsection.

Let's verify that the Poisson distribution defines a valid probability mass function by checking that the total probability sums to 1:

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}.$$

Using the Taylor expansion for the exponential function:

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda},$$

we conclude that:

$$e^{-\lambda} \cdot e^{\lambda} = 1.$$

Example. At *Probability Airlines*, the customer service call center receives incoming calls at an average rate of $\lambda = 3$ per minute. This situation can be modeled using the Poisson distribution.

For example, the probability of exactly 5 calls arriving in a one-minute interval is:

$$P(X = 5) = \frac{3^5}{5!} \cdot e^{-3} = \frac{243}{120} \cdot e^{-3} \approx 0.0136.$$

Now, let X and Y be two independent random variables that follow Poisson distributions with parameters λ and μ, respectively:

$$X \sim \text{Poisson}(\lambda), \quad Y \sim \text{Poisson}(\mu).$$

We are interested in the distribution of the sum $Z = X + Y$. The probability mass function of Z is given by the convolution formula:

$$P(Z = n) = \sum_{k=0}^{n} P(X = k \text{ and } Y = n - k) = \sum_{k=0}^{n} P(X = k) \cdot P(Y = n - k) = \sum_{k=0}^{n} \frac{\lambda^k}{k!} \cdot e^{-\lambda} \cdot \frac{\mu^{n-k}}{(n-k)!} \cdot e^{-\mu} =$$

$$\frac{e^{-(\lambda+\mu)}}{n!} \cdot \sum_{k=0}^{n} \binom{n}{k} \lambda^k \mu^{n-k} = \frac{(\lambda + \mu)^n}{n!} e^{-(\lambda+\mu)}.$$

This final expression is the probability mass function of a Poisson distribution with parameter $\lambda + \mu$. Thus, we conclude:

$$X + Y \sim \text{Poisson}(\lambda + \mu).$$

Remark. If the number of events occurring in a time interval follows a Poisson distribution with parameter $\lambda > 0$, then the time between consecutive events follows an Exponential distribution with the same $\lambda > 0$.

Application to Bitcoin Mining

The arrival of new blocks in the Bitcoin network is commonly modeled as a Poisson process with a constant average rate (approximately one block every 10 minutes). Consequently, the time between successive blocks (i.e., the block mining times) is modeled as an exponential random variable. This framework allows for the analysis of network behavior, miner strategies, and expected rewards based on stochastic properties of exponential waiting times.

Optional: Computing the Expected Value of a Poisson Random Variable

We now compute the expected value of a Poisson-distributed random variable X, with parameter $\lambda > 0$ The expected value is given by

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} P(X = k) \cdot k = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda}.$$

To evaluate this sum, we employ an idea similar to one used in Lecture 10, where we computed the expected number of tosses before the first Heads. Define the function

$$F(\lambda) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda},$$

which, as we have observed, satisfies $F(\lambda) = 1$ for all λ, since this is the total probability.

We now compute the derivative $F'(\lambda)$ using the product rule:

$$F'(\lambda) = \left(e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \right)' = -e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} + e^{-\lambda} \sum_{k=0}^{\infty} \left(\frac{\lambda^k}{k!} \right)'.$$

We simplify the second term:

$$\left(\frac{\lambda^k}{k!} \right)' = k \cdot \frac{\lambda^{k-1}}{k!} \quad \text{for } k \geq 1.$$

So we get:

$$F'(\lambda) = -e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} + e^{-\lambda} \sum_{k=1}^{\infty} k \cdot \frac{\lambda^{k-1}}{k!}.$$

Now, multiply both sides of the latter expression by λ:

$$\lambda F'(\lambda) = -\lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} + e^{-\lambda} \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!}.$$

Recognizing that the last term is exactly $\mathbb{E}(X)$, and using that the first sum equals 1, we obtain:

$$\lambda F'(\lambda) = -\lambda + \mathbb{E}(X).$$

Since $F'(\lambda) = 0$, we conclude that $\mathbb{E}(X) = \lambda$.

Remark. This elegant trick is based on differentiating a generating function and recognizing how derivatives “pull down” factors of k in the expected value sum. It is a useful strategy for many other distributions as well.

Lecture 16
Conditional Distribution and Expectation

Consider two random variables, X and Y, with joint probability function $p(x, y)$. The conditional probability $P(X = x \mid Y = y)$ represents the likelihood that X takes the value x, given that we have observed $Y = y$. This is analogous to conditional probability for events: just as

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

updates our belief about event A after learning that event B occurred, the conditional probability of X given $X = x$ tells us how to update our knowledge about the value of Y once we know the value of X.

Definition. Let X and Y be two random variables.

- If X and Y are discrete with joint PMF $p(x, y)$, the **conditional PMF of Y given $X = x$** is defined as
$$p(y \mid x) = \frac{p(x, y)}{p_X(x)} = \frac{p(x, y)}{\sum\limits_{y' \in Y} p(x, y')}.$$

- If X and Y are continuous with joint PDF $f(x, y)$, the **conditional PDF of Y given $X = x$** is defined as
$$f_{Y|X}(x, y) = \frac{f(x, y)}{f_X(x)} = \frac{f(x, y)}{\int\limits_{-\infty}^{\infty} f(x, y')\, dy'}.$$

Example. Consider the continuous random variables (X, Y) with joint probability density function

$$f(x, y) = \begin{cases} \frac{3}{2}(x^2 + y^2), & \text{if } 0 \leq x, y \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

We aim to compute the conditional probability density function of Y given X.
Step 1. Compute the marginal density of X:

$$f_X(x) = \int\limits_0^1 \frac{3}{2}(x^2 + y^2)\, dy = \frac{3}{2}\left(x^2 y + \frac{1}{3}y^3\right)\Bigg|_{y=0}^{1} = \frac{3}{2}\left(x^2 + \frac{1}{3}\right).$$

Step 2. Compute the conditional density of Y given $X = x$:

$$f_{Y|X}(x, y) = \frac{f(x, y)}{f_X(x)} = \frac{\frac{3}{2}(x^2 + y^2)}{\frac{3}{2}\left(x^2 + \frac{1}{3}\right)} = \frac{x^2 + y^2}{x^2 + \frac{1}{3}}.$$

Now we can specialize $f_{Y|X}(x, y)$ to concrete values. For instance, suppose $X = 0.4$. Then the conditional PDF becomes

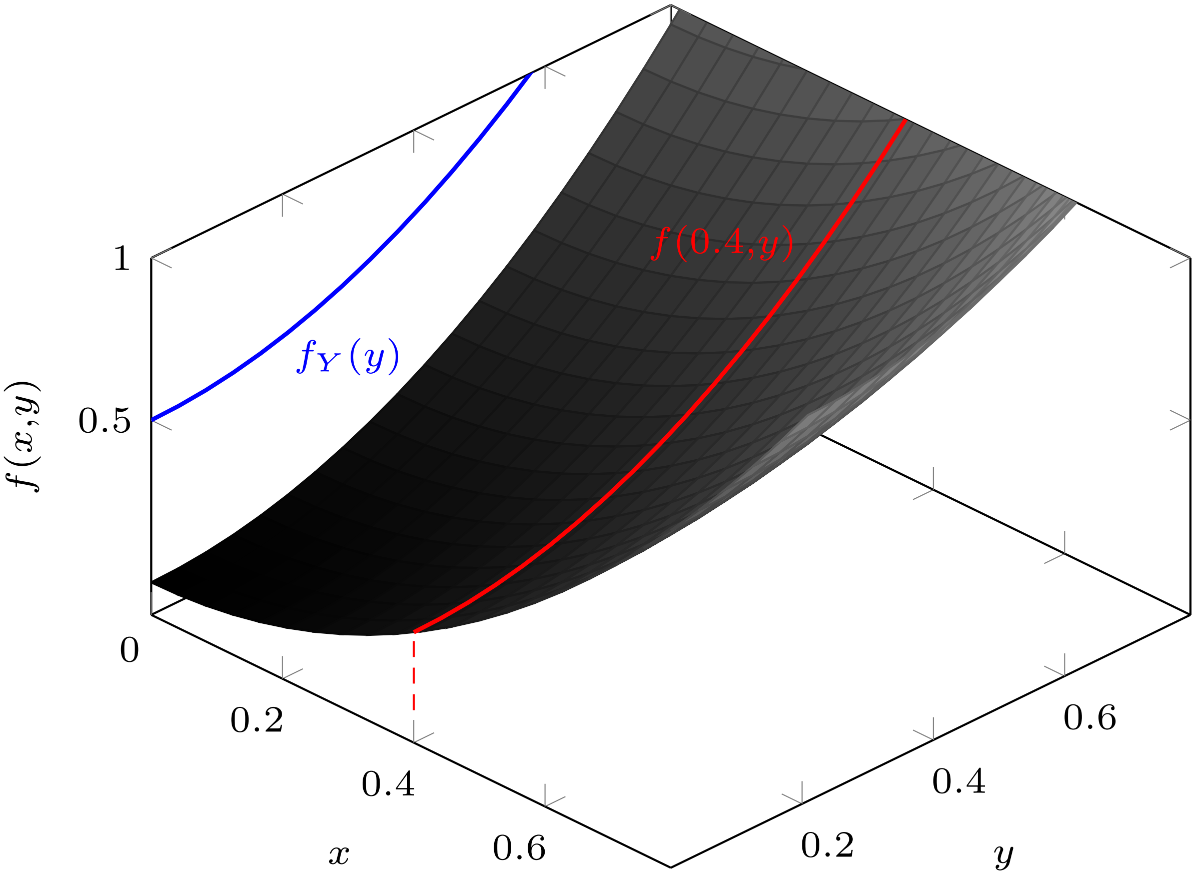
$$f_{Y|X=0.4}(y) = \frac{f(0.4, y)}{f_X(0, 4)} = \frac{0.16 + y^2}{0.16 + \frac{1}{3}}.$$

It is worth verifying that

$$\int\limits_0^1 \frac{0.16 + y^2}{0.16 + \frac{1}{3}}\, dy = \left(\frac{0.16y + \frac{y^3}{3}}{0.16 + \frac{1}{3}}\right)\Bigg|_0^1 = 1,$$

confirming the validity of $f_{Y|X=0.4}(y)$ as a proper probability density function.

The joint distribution $f(x, y)$ can be visualized as a surface in 3D space over the (x, y)-plane. The conditional distribution $f_{Y|X=a}(y)$ corresponds to a vertical slice of this surface at $x = a$, normalized to integrate to 1, while the marginal density $f_Y(y)$ is obtained by collapsing the surface along the x-axis via integrating out x:



Conditional Expectation

Let X and Y be two random variables with a known joint distribution.

Suppose we fix a value x_0 of the random variable X. Conditioning on $X = x_0$ defines a new probability distribution for Y, and therefore we may compute quantities such as the expectation or variance of Y under this conditional distribution.

Definition. The **conditional expectation of** Y **given** $X = x$ is the expected value of Y under the conditional distribution of Y given $X = x$. This is written as

$$\mathbb{E}(Y \mid X = x).$$

- If X and Y are discrete random variables with joint probability mass function $p(x, y)$, then the conditional expectation is given by

  $$\mathbb{E}(Y \mid X = x) = \sum_y y p_{Y|X}(x, y) = \frac{1}{p_X(x)} \sum_y y \cdot p(x, y),$$

  where $p_{Y|X}(x, y)$ is the conditional PMF of Y given $X = x$.

- If X and Y are continuous random variables with joint density $f(x, y)$, then the conditional expectation is given by

  $$\mathbb{E}(Y \mid X) = \int_{-\infty}^{\infty} y f_{Y|X}(x, y)\, dy = \frac{1}{f_X(x)} \int_{-\infty}^{\infty} y f(x, y)\, dy,$$

  where $f_{Y|X}(x, y) = \frac{f(x,y)}{f_X(x)}$ is the conditional density of Y given $X = x$.

Remark. Our original joint PDF was a function of both x and y. By integrating over y, the conditional expectation $\mathbb{E}[Y \mid X = x]$ becomes a function of x alone.

One can interpret this as an ordinary expectation normalized by the marginal density of X. More precisely, the conditional expectation is a normalized version of the joint expectation, with normalization provided by the marginal density $f_X(x)$.

Example. Consider two random variables X and Y with a joint probability density function given by:

$$f(x, y) = \begin{cases} 2, & 0 < x < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

To compute the marginal PDF of X, we integrate out y from the joint PDF:

$$f_X(x) = \int\limits_x^1 2\,dy = 2 - 2x, \quad 0 < x < 1.$$

Likewise, the marginal PDF of Y is obtained by integrating out x:

$$f_Y(y) = \int\limits_0^y 2\,dx = 2y, \quad 0 < y < 1.$$

We now compute the conditional expectation $\mathbb{E}(Y \mid X = x)$. Since the support of Y given $X = x$ is $y \in (x, 1)$, we have:

$$\mathbb{E}(Y \mid X = x) = \frac{1}{f_X(x)} \int\limits_x^1 y f(x, y)\,dy = \frac{1}{2 - 2x} \int\limits_x^1 2y\,dy = \frac{y^2}{2 - 2x}\bigg|_{y=x}^1 = \frac{1 - x^2}{2(1 - x)} = \frac{1 + x}{2}.$$

Next, we compute $\mathbb{E}(X \mid Y = y)$. Since $x \in (0, y)$, we have:

$$\mathbb{E}(X \mid Y = y) = \frac{1}{f_Y(y)} \int\limits_0^y x f(x, y)\,dx = \frac{1}{2y} \int\limits_0^y 2x f(x, y)\,dx = \frac{x^2}{2y}\bigg|_{x=0}^y = \frac{y^2}{2y} = 0.5y.$$

Lecture 17

Covariance and Correlation

We have spent some time working with pairs of random variables, exploring joint distributions and how two variables might interact. A natural question now arises: how can we formally capture and quantify the nature of the relationship between two random variables?

To motivate this idea, let us consider two practical examples in which several variables are interrelated. These real-world scenarios highlight the intuition that when one quantity changes, another may tend to increase or decrease in response. Our goal is to move beyond qualitative observations and develop a mathematical framework for such interactions.

Laptop Performance and Battery Life

In this scenario, we focus on the quadruple (C, R, P, B):

- C represents the CPU clock speed (in gigahertz).
- R represents the amount of RAM (in gigabytes).
- P represents the storage type (e.g., HDD or SSD) and capacity (in gigabytes).
- B represents the battery life (in hours).

It is generally expected that as C, R, and P increase, the corresponding battery life B tends to decrease. This is due to the fact that higher CPU clock speeds, larger RAM, and larger storage capacities typically require more power, resulting in reduced battery life. However, advancements in technology and power optimization techniques can influence these relationships to some extent.

Impact of Physical Exercise on Health

Here we examine the quadruple (A, W, E, H):

- A represents the age of individuals.
- W represents the frequency of physical exercise.
- E represents the quality of eating habits.
- H represents the overall health condition.

It is generally expected that individuals who engage in regular exercise (higher W) tend to have better overall health (higher H), regardless of age (A) and eating habits (E). While individual cases may differ due to various factors, a positive association between exercise and health is commonly observed.

Our next step is to develop formal tools to describe such patterns of association between random variables: this leads us to the notions of *covariance* and *correlation*.

The Formalities

Let X and Y be two random variables with means $\mathbb{E}(X) = \mu_1$ and $\mathbb{E}(Y) = \mu_2$ respectively.

Definition. The **covariance** of X and Y is defined by

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_1)(Y - \mu_2)].$$

This expression measures how the variables X and Y vary in tandem relative to their respective means. The quantity $(X - \mu_1)(Y - \mu_2)$ is positive if both variables are simultaneously above or below their averages, and negative if one is above while the other is below.

- A positive covariance suggests that larger values of X tend to be paired with larger values of Y, and smaller with smaller.
- A negative covariance suggests that larger values of X tend to be paired with smaller values of Y, and vice versa.
- A covariance near zero suggests little or no linear relationship between X and Y.

Just like variance, covariance can be rewritten in a more computationally efficient form. Using linearity of expectation, we obtain

$$\begin{aligned}&\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_1)(Y - \mu_2)] = \mathbb{E}[XY - X \cdot \mu_2 - Y \cdot \mu_1 + \mu_1\mu_2] = \\ &\mathbb{E}(XY) - \mu_2\mathbb{E}(X) - \mu_1\mathbb{E}(Y) + \mu_1\mu_2 = \mathbb{E}(XY) - \mu_2\mu_1 - \mu_1\mu_2 + \mu_1\mu_2 = \\ &\mathbb{E}(XY) - \mu_1\mu_2.\end{aligned}$$

In other words, we can compute covariance as

$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Remark. The covariance between a random variable X and itself is given by

$$\mathrm{Cov}(X, X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathrm{Var}(X),$$

which is precisely the variance of X.

Example. We now illustrate the computation of covariance in two settings involving pairs of random variables.

1. Consider flipping a fair coin three times. Define X as the number of heads in the first two flips and Y as the total number of heads in all three flips. The marginal distributions of X and Y yield:

$$\mathbb{E}(X) = \frac{2}{8}\cdot 0 + \frac{4}{8}\cdot 1 + \frac{2}{8}\cdot 2 = 1, \qquad \mathbb{E}(Y) = \frac{1}{8}\cdot 0 + \frac{3}{8}\cdot 1 + \frac{3}{8}\cdot 2 + \frac{1}{8}\cdot 3 = \frac{3}{2}.$$

Next we evaluate

$$\mathbb{E}(XY) = \frac{2}{8}\cdot 1\cdot 1 + \frac{2}{8}\cdot 1\cdot 2 + \frac{1}{8}\cdot 2\cdot 2 + \frac{1}{8}\cdot 2\cdot 3 = 2,$$

so that

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 2 - 1\cdot\frac{3}{2} = \frac{1}{2}.$$

2. Now let X and Y be continuous random variables with joint probability density function

$$f(x,y) = \begin{cases} x+y & \text{for } 0 < x, y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

We begin by computing the expected values:

$$\mathbb{E}(X) = \int_0^1\int_0^1 x(x+y)\,dy\,dx = \int_0^1 \left(x^2y + \frac{1}{2}xy^2\right)\Bigg|_{y=0}^1 dx = \int_0^1 \left(x^2 + \frac{x}{2}\right) dx = \left(\frac{x^3}{3} + \frac{x^2}{4}\right)\Bigg|_0^1 = \frac{7}{12},$$

and similarly,

$$\mathbb{E}(Y) = \int_0^1\int_0^1 y(x+y)\,dx\,dy = \int_0^1 \left(xy^2 + \frac{1}{2}yx^2\right)\Bigg|_{x=0}^1 dy = \int_0^1 \left(y^2 + \frac{y}{2}\right) dy = \left(\frac{y^3}{3} + \frac{y^2}{4}\right)\Bigg|_0^1 = \frac{7}{12}.$$

The cross term is evaluated as follows:

$$\mathbb{E}(XY) = \int_0^1\int_0^1 xy(x+y)\,dx\,dy = \int_0^1\int_0^1 (x^2y + xy^2)\,dx\,dy = \int_0^1 \left(\frac{x^3y}{3} + \frac{x^2y^2}{2}\right)\Bigg|_{x=0}^1 dy =$$
$$\int_0^1 \left(\frac{y}{3} + \frac{y^2}{2}\right) dy = \left(\frac{y^2}{6} + \frac{y^3}{6}\right)\Bigg|_0^1 = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

Finally, we find the covariance:

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \frac{1}{3} - \left(\frac{7}{12}\right)^2 = -\frac{1}{144}.$$

Correlation as a Normalized Measure of Linear Dependence

When two random variables X and Y exhibit a high covariance, this may indicate a strong linear relationship. However, it might also arise simply because both variables have large variances. To better assess the strength of the linear relationship between X and Y, we normalize the covariance by the product of their standard deviations.

Definition. The **correlation** between two random variables X and Y is defined as

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y},$$

where $\sigma_X = \sqrt{\mathrm{Var}(X)}$ and $\sigma_Y = \sqrt{\mathrm{Var}(Y)}$ are the respective standard deviations of X and Y.

Remark. This definition is valid only when both variables have positive variance, i.e., $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$. If either variable is constant, its variance is zero, and the correlation is undefined—since it makes no sense to speak of variation with respect to a constant.

As a special case, the correlation of a variable with itself is always equal to 1:

$$\rho(X,X) = \frac{\mathrm{Cov}(X,X)}{\sigma_X^2} = \frac{\mathrm{Var}(X)}{\mathrm{Var}(X)} = 1.$$

Example. We continue with the two examples considered earlier in this lecture. Having already computed the covariances, we now calculate the variances of each random variable and use them to determine the corresponding correlation coefficients.

1. Recall the values of the random variables X and Y, along with their joint distribution. We compute

$$\mathbb{E}(X^2) = \frac{2}{8} \cdot 0^2 + \frac{4}{8} \cdot 1^2 + \frac{2}{8} \cdot 2^2 = \frac{3}{2},$$
$$\mathbb{E}(Y^2) = \frac{1}{8} \cdot 0^2 + \frac{3}{8} \cdot 1^2 + \frac{3}{8} \cdot 2^2 + \frac{1}{8} \cdot 3^2 = 3.$$

   Using these, the variances are

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \frac{3}{2} - 1 = 0.5,$$
$$\mathrm{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}^2(Y) = 3 - \frac{9}{4} = 0.75.$$

   The correlation coefficient is

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{0.5}{\sqrt{0.5} \cdot \sqrt{0.75}} \approx 0.816.$$

2. In the case of two continuous random variables with a joint PDF

$$f(x,y) = \begin{cases} x + y & \text{for } 0 < x, y < 1, \\ 0 & \text{otherwise,} \end{cases}$$

   we have

$$\mathbb{E}(X^2) = \int_0^1 \int_0^1 x^2(x+y)\,dy\,dx = \int_0^1 \left(x^3 y + \frac{x^2 y^2}{2} \right) \Big|_{y=0}^{1} dx = \int_0^1 \left(x^3 + \frac{x^2}{2} \right) dx =$$
$$\left(\frac{x^4}{4} + \frac{x^3}{6} \right) \Big|_0^1 = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}.$$

Since we previously found $\mathbb{E}(X) = \frac{7}{12}$, we get

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \frac{5}{12} - \frac{49}{144} = \frac{11}{144}.$$

By symmetry, $\mathrm{Var}(Y) = \frac{11}{144}$ as well. With covariance $\mathrm{Cov}(X,Y) = -\frac{1}{144}$, we obtain the correlation coefficient of

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{-\frac{1}{144}}{\sqrt{\frac{11}{144}} \cdot \sqrt{\frac{11}{144}}} = -\frac{1}{11}.$$

Theorem. Let X and Y be random variables with positive variances. Then the correlation coefficient $\rho(X,Y)$ satisfies the inequality

$$-1 \leq \rho(X,Y) \leq 1.$$

Optional: Proof of $|\rho(X,Y)| \leq 1$ via the Cauchy–Schwarz Inequality

We give a short proof of the inequality $|\rho(X,Y)| \leq 1$, which bounds the correlation coefficient between two random variables. The argument is a direct application of the Cauchy–Schwarz inequality.

Proof. We begin by observing two useful facts:

- Variance is translation-invariant: $\mathrm{Var}(X) = \mathrm{Var}(X - a)$ for any constant $a \in \mathbb{R}$.
- Covariance is translation-invariant in each argument: $\mathrm{Cov}(X-a, Y-b) = \mathrm{Cov}(X,Y)$ for any $a, b \in \mathbb{R}$.

Therefore, by subtracting the respective means, if necessary, we can assume that $\mathbb{E}(X) = \mathbb{E}(Y) = 0$. In this case, the variances reduce to

$$\mathrm{Var}(X) = \mathbb{E}(X^2), \quad \mathrm{Var}(Y) = \mathbb{E}(Y^2),$$

and the covariance becomes $\mathrm{Cov}(X,Y) = \mathbb{E}(XY)$.

Now consider the family of random variables $Z_t = X + tY$, where $t \in \mathbb{R}$. Since $Z_t^2 \geq 0$, we have

$$\mathbb{E}(Z_t^2) = \mathbb{E}((X+tY)^2) = \mathbb{E}(X^2) + 2t\,\mathbb{E}(XY) + t^2\,\mathbb{E}(Y^2) \geq 0$$

for all real t. This expression is a quadratic polynomial in t with leading coefficient $\mathbb{E}(Y^2) > 0$, so its discriminant must be non-positive:

$$D = 4\mathbb{E}(XY)^2 - 4\mathbb{E}(X^2)\mathbb{E}(Y^2) \leq 0.$$

Dividing both sides by 4, we get

$$\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2),$$

or equivalently,

$$\left| \frac{\mathbb{E}(XY)}{\sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}} \right| \leq 1.$$

Recalling that $\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}(XY)}{\sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}}$ under our zero-mean assumption, we conclude:

$$|\rho(X,Y)| \leq 1.$$

□

Remark. This inequality is a special case of the Cauchy–Schwarz inequality, which states that for any two vectors u and v in an inner product space,

$$|\langle u, v \rangle| \leq \|u\| \cdot \|v\|.$$

In our context, we treat the centered random variables $X - \mathbb{E}(X)$ and $Y - \mathbb{E}(Y)$ as vectors, define the inner product by $\langle X, Y \rangle := \mathbb{E}(XY)$, and the norm by $\|X\| = \sqrt{\mathbb{E}(X^2)} = \sigma_X$. Then the Cauchy–Schwarz inequality gives

$$|\mathbb{E}(XY)| \leq \sigma_X \sigma_Y,$$

which is equivalent to $|\rho(X,Y)| \leq 1$.

Lecture 18
Independence of Two Random Variables

Recall our earlier definition of independence for events A and B: the events are said to be *independent* if

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

or equivalently,

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

This notion captures the idea that knowing A occurred does not affect the probability of B.

We now extend this idea to random variables. Informally, two random variables X and Y are independent if knowing the value of one provides no information about the distribution of the other. Formally, this is defined in terms of conditional and marginal distributions.

Definition. Two **random variables** X and Y are said to be **independent** if the conditional distribution of Y given $X = x$ is equal to the marginal distribution of Y for all values x:

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

Using the definition of conditional density,

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

we conclude that independence implies

$$\frac{f(x, y)}{f_X(x)} = f_Y(y) \quad \Rightarrow \quad f(x, y) = f_X(x) \cdot f_Y(y).$$

In other words, two random variables are independent if and only if their joint probability density function factors as the product of their marginal densities, mirroring the formula $P(A \cap B) = P(A)P(B)$ from event-level probability. where probabilities are replaced by densities.

Example. Consider two random variables X and Y with joint probability density function

$$f(x, y) = 4xy,$$

defined on the unit square $0 \le x \le 1$, $0 \le y \le 1$. We will check if X and Y are independent.

Step 1. Compute the marginal distribution of X by integrating over y:

$$f_X(x) = \int_0^1 4xy\, dy = 4x \int_0^1 y\, dy = 4x \cdot \frac{1}{2} = 2x.$$

Step 2. Compute the conditional distribution of Y given $X = x$:

$$f_{Y|X}(x, y) = \frac{f(x, y)}{f_X(x)} = \frac{4xy}{2x} = 2y.$$

Note that the conditional density $f_{Y|X}(x, y) = 2y$ is independent of x.

Step 3. Compute the marginal distribution of Y:

$$f_Y(y) = \int_0^1 4xy\, dx = 4y \int_0^1 x\, dx = 4y \cdot \frac{1}{2} = 2y.$$

We observe that

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

which indicates that knowing the value of X does not affect the distribution of Y. Therefore, X and Y are independent.

Remark. An equivalent condition for the independence of continuous random variables X and Y is that their joint probability density function factors as

$$f(x,y) = g(x)h(y),$$

where $g(x)$ is a function of x alone and $h(y)$ is a function of y alone. The factorization of a joint density into marginal components is not unique. In the example above, the joint density $f(x,y) = 4xy$ can be written as $2x \cdot 2y$, or as $x \cdot 4y$, and so on. The key point is that each factor involves only one of the variables.

This criterion emphasizes that the joint behavior of X and Y arises solely from their individual distributions. That is, knowledge of one variable provides no information about the other—precisely the hallmark of independence.

Proposition. If the random variables X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$.

Proof (for continuous random variables). We present the proof for continuous random variables; the discrete case follows analogously.

To compute $\mathbb{E}(X)$, we integrate over the joint distribution:

$$\mathbb{E}(X) = \iint x f(x,y)\,dy\,dx = \int x \left(\int f(x,y)\,dy \right) dx = \int x f_X(x)\,dx,$$

where $f_X(x)$ is the marginal density of X.

Now assume X and Y are independent. Then the joint density factorizes:

$$f(x,y) = f_X(x) f_Y(y),$$

and so the expectation of the product is

$$\mathbb{E}(XY) = \iint xy f(x,y)\,dy\,dx = \iint xy f_X(x) f_Y(y)\,dy\,dx = \left(\int x f_X(x)\,dx \right) \left(\int y f_Y(y)\,dy \right) = \mathbb{E}(X)\mathbb{E}(Y).$$

Substituting into the covariance formula, we get:

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0.$$

□

Remark. The converse of the proposition is generally false: zero covariance does not imply independence. For a counterexample, see Problem 7 in Final Exam, Part I.