Statistical Inference

Lecture 1: Probability Theory

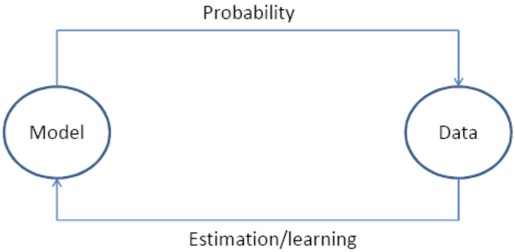
MING GAO

DASE @ ECNU
(for course related communications)
mgao@dase.ecnu.edu.cn

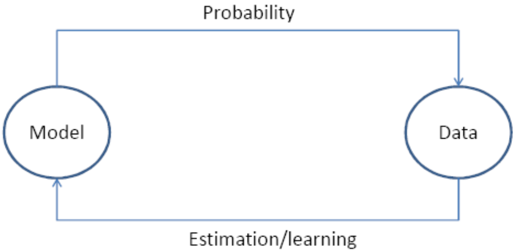
Mar. 1, 2018

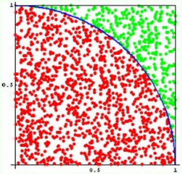
Outline

Introduction



Introduction





Probability as a mathematical framework for:

- reasoning about uncertainty
- deriving approaches to inference problems

Experiment and sample space

Experiment

An **experiment** is a procedure that yields one of a given set of possible outcomes.

Experiment and sample space

Experiment

An **experiment** is a procedure that yields one of a given set of possible outcomes.

Sample space

The **sample space**, denoted as Ω, of the experiment is the set of possible outcomes.

Experiment and sample space

Experiment

An **experiment** is a procedure that yields one of a given set of possible outcomes.

Sample space

The **sample space**, denoted as Ω, of the experiment is the set of possible outcomes.

Example

- Roll a die one time, $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- We toss a coin twice (Head = H, Tail = T), $\Omega = \{HH, HT, TH, TT\}$.

Experiment and sample space

Experiment

An **experiment** is a procedure that yields one of a given set of possible outcomes.

Sample space

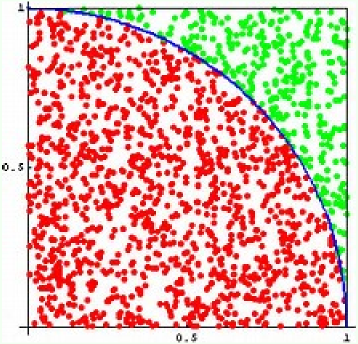
The **sample space**, denoted as Ω, of the experiment is the set of possible outcomes.

Example

- Roll a die one time, $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- We toss a coin twice (Head = H, Tail = T), $\Omega = \{HH, HT, TH, TT\}$.

- "List" (set) of possible outcomes
- List must be:
  1. Mutually exclusive
  2. Collectively exhaustive
- Art: to be at the "right" granularity

Continuous sample space



For this case, sample space $\Omega = \{(x, y) | 0 \leq x, y \leq 1\}$. Note that the sample space is infinite and uncountable.

In this course, we consider the **countable sample spaces** and **uncountable sample spaces**. Thus, we call the learning content to be the discrete probability and continuous probability, respectively.

Event and set operators

An **event**, represented as a set, is a subset of the sample space.

Event and set operators

An **event**, represented as a set, is a subset of the sample space.

Example

- Toss at least one head $B = \{HH, HT, TH\} \subset \Omega$;
- Toss at least three head $C = \emptyset \subset \Omega$.
- There are $2^{|\Omega|}$ events for an experiments;
- Events therefore have all set operations.

Event and set operators

An **event**, represented as a set, is a subset of the sample space.

Example

- Toss at least one head $B = \{HH, HT, TH\} \subset \Omega$;
- Toss at least three head $C = \emptyset \subset \Omega$.
- There are $2^{|\Omega|}$ events for an experiments;
- Events therefore have all set operations.

Operators

- **Containment:** $A \subset B \Leftrightarrow x \in A \Rightarrow x \in B$;
- **Union:** $A \cup B = \{x | x \in A \text{ or } x \in B\}$;
- **Intersection:** $A \cap B = \{x | x \in A \text{ and } x \in B\}$;
- **Difference**: $A - B = \{x : x \in A \wedge x \notin B\}$;
- **Complement:** $A^c = \{x | x \notin A\}$;

Set identities

Table of set identities

equivalence	name
$A \cap U = A$ $A \cup \emptyset = A$	Identity laws
$A \cap A = A$ $A \cup A = A$	Idempotent laws
$(A \cap B) \cap C = A \cap (B \cap C)$ $(A \cup B) \cup C = A \cup (B \cup C)$	Associative laws
$A \cup (A \cap B) = A$ $A \cap (A \cup B) = A$	Absorption laws
$A \cap \overline{A} = \emptyset$ $A \cup \overline{A} = U$	Complement laws

Logical equivalence Cont'd

Table of set identities

equivalence	name
$A \cup U = U$ $A \cap \emptyset = \emptyset$	Domination laws
$A \cap B = B \cap A$ $A \cup B = B \cup A$	Commutative laws
$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$	Distributive laws
$\overline{(A \cap B)} = \overline{A} \cup \overline{B}$ $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$	De Morgan's laws

Extensions of set operators

If $A_1, A_2, \cdots, A_n, \cdots$ is a collection of events, all defined on a sample space Ω, then

- **Union:** $\bigcup_{i=1}^{\infty} A_i = \{x \in \Omega | x \in A_i \text{ for some } i\}$;
- **Intersection:** $\bigcap_{i=1}^{\infty} A_i = \{x \in \Omega | x \in A_i \text{ for all } i\}$;

Example

Let $\Omega = (0, 1]$ and define $A_i = [\frac{1}{i}, 1]$. then

- $\bigcup_{i=1}^{\infty} A_i = \lim_{n\to+\infty} \bigcup_{i=1}^{n} [\frac{1}{i}, 1] = \lim_{n\to+\infty} [\frac{1}{n}, 1] = (0, 1]$;
- $\bigcap_{i=1}^{\infty} A_i = \lim_{n\to+\infty} \bigcap_{i=1}^{n} [\frac{1}{i}, 1] = \lim_{n\to+\infty} [1, 1] = \{1\}$;

Extensions of set operators Cont'd

If Γ is an uncountable collection of events, all defined on a sample space Ω, then

- **Union:** $\bigcup_{a \in \Gamma} A_i = \{x \in \Omega | x \in A_a \text{ for some } a\}$;
- **Intersection:** $\bigcap_{a \in \Gamma} A_i = \{x \in \Omega | x \in A_a \text{ for all } a\}$;

Example

Let $\Gamma = R^+$ and define $A_a = (0, a]$. then

- $\bigcup_{a \in \Gamma} A_a = \lim_{a \to +\infty} \bigcup_{i=1}^{a} (0, a] = (0, +\infty]$;
- $\bigcap_{a \in \Gamma} A_a = \lim_{a \to +\infty} \bigcap_{i=1}^{a} (0, a] = \emptyset$;

Disjoint and partition

- Two events A and B are disjoint (or mutually exclusive) if $A \cap B = \emptyset$. The events $A_1, A_2, \cdots, A_n, \cdots$ are pairwise disjoint (or mutually exclusive) if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Disjoint and partition

- Two events A and B are disjoint (or mutually exclusive) if $A \cap B = \emptyset$. The events $A_1, A_2, \cdots, A_n, \cdots$ are pairwise disjoint (or mutually exclusive) if $A_i \cap A_j = \emptyset$ for all $i \neq j$.
- If $A_1, A_2, \cdots, A_n, \cdots$ are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = \Omega$, then the collection $A_1, A_2, \cdots, A_n, \cdots$ forms a partition of Ω.

Disjoint and partition

- Two events A and B are disjoint (or mutually exclusive) if $A \cap B = \emptyset$. The events $A_1, A_2, \cdots, A_n, \cdots$ are pairwise disjoint (or mutually exclusive) if $A_i \cap A_j = \emptyset$ for all $i \neq j$.
- If $A_1, A_2, \cdots, A_n, \cdots$ are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = \Omega$, then the collection $A_1, A_2, \cdots, A_n, \cdots$ forms a partition of Ω.
- The collection $A_i = [i, i+1)$ for $i \in N$ consists of pairwise disjoint sets. Furthermore, we have $\bigcup_{i=1}^{\infty} A_i = [0, +\infty)$.

Disjoint and partition

- Two events A and B are disjoint (or mutually exclusive) if $A \cap B = \emptyset$. The events $A_1, A_2, \cdots, A_n, \cdots$ are pairwise disjoint (or mutually exclusive) if $A_i \cap A_j = \emptyset$ for all $i \neq j$.
- If $A_1, A_2, \cdots, A_n, \cdots$ are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = \Omega$, then the collection $A_1, A_2, \cdots, A_n, \cdots$ forms a partition of Ω.

- The collection $A_i = [i, i+1)$ for $i \in N$ consists of pairwise disjoint sets. Furthermore, we have $\bigcup_{i=1}^{\infty} A_i = [0, +\infty)$.
- The collection of sets $A_i = [i, i+1)$ for $i \in N$ forms a partition of $[0, +\infty)$.

Sigma Algebra

A collection of subsets of Ω is called a sigma algebra (or Borel field), denoted as $\mathcal{B}$, if it satisfies the following three properties

- $\emptyset \in \mathcal{B}$;

Sigma Algebra

A collection of subsets of Ω is called a sigma algebra (or Borel field), denoted as $\mathcal{B}$, if it satisfies the following three properties

- $\emptyset \in \mathcal{B}$;
- If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$;

Sigma Algebra

A collection of subsets of Ω is called a sigma algebra (or Borel field), denoted as $\mathcal{B}$, if it satisfies the following three properties

- $\emptyset \in \mathcal{B}$;
- If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$;
- If $A_1, A_2, \cdots \in \mathcal{B}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$ ($\mathcal{B}$ is closed under countable unions).

Sigma Algebra

A collection of subsets of Ω is called a sigma algebra (or Borel field), denoted as $\mathcal{B}$, if it satisfies the following three properties

- $\emptyset \in \mathcal{B}$;
- If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$;
- If $A_1, A_2, \cdots \in \mathcal{B}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$ ($\mathcal{B}$ is closed under countable unions).

In addition, from DeMorgan's Laws it follows that $\mathcal{B}$ is closed under countable intersection, i.e.,

$$\Big(\bigcup_{i=1}^{\infty} A_i\Big)^c = \bigcap_{i=1}^{\infty} A_i^c \in \mathcal{B}.$$

Example

If Ω is finite or countable, the sigma algebra of Ω can be defined as

$$\mathcal{B} = \{A | A \subseteq \Omega\}.$$

- If Ω has n elements, there are 2^n sets in $\mathcal{B}$;

Example

If Ω is finite or countable, the sigma algebra of Ω can be defined as

$$\mathcal{B} = \{A | A \subseteq \Omega\}.$$

- If Ω has n elements, there are 2^n sets in $\mathcal{B}$;
- If Ω is countable infinite, the cardinality of $\mathcal{B}$ is $\aleph_0$;

Example

If Ω is finite or countable, the sigma algebra of Ω can be defined as

$$\mathcal{B} = \{A | A \subseteq \Omega\}.$$

- If Ω has n elements, there are 2^n sets in $\mathcal{B}$;
- If Ω is countable infinite, the cardinality of $\mathcal{B}$ is $\aleph_0$;
- Let $\Omega = (-\infty, +\infty)$, then $\mathcal{B}$ is chosen to contain all sets of the form

  $$[a, b], (a, b], (a, b) \text{ and } [a, b)$$

  for all real numbers a and b.

Probability function

Definition

Given a sample space Ω and an associated sigma algebra $\mathcal{B}$, a probability function is a function P with domain $\mathcal{B}$ that satisfies

Probability function

Definition

Given a sample space Ω and an associated sigma algebra $\mathcal{B}$, a probability function is a function P with domain $\mathcal{B}$ that satisfies

- **Nonnegativity:** $P(A) \geq 0$ for all $A \in \mathcal{B}$

Probability function

Definition

Given a sample space Ω and an associated sigma algebra $\mathcal{B}$, a probability function is a function P with domain $\mathcal{B}$ that satisfies

- **Nonnegativity:** $P(A) \geq 0$ for all $A \in \mathcal{B}$
- **Normalization:** $P(\Omega) = 1$ and $P(\emptyset) = 0$;

Probability function

Definition

Given a sample space Ω and an associated sigma algebra $\mathcal{B}$, a probability function is a function P with domain $\mathcal{B}$ that satisfies

- **Nonnegativity:** $P(A) \geq 0$ for all $A \in \mathcal{B}$
- **Normalization:** $P(\Omega) = 1$ and $P(\emptyset) = 0$;
- **Additivity:** If $A_1, A_2, \dots \in \mathcal{B}$ are pairwise disjoint, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Probability function

Definition

Given a sample space Ω and an associated sigma algebra $\mathcal{B}$, a probability function is a function P with domain $\mathcal{B}$ that satisfies

- **Nonnegativity:** $P(A) \geq 0$ for all $A \in \mathcal{B}$
- **Normalization:** $P(\Omega) = 1$ and $P(\emptyset) = 0$;
- **Additivity:** If $A_1, A_2, \dots \in \mathcal{B}$ are pairwise disjoint, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Note that

- The three properties given in the definition are usually referred to as the Axioms of probability (or the Kolmogorov Axioms).

Probability function

Definition

Given a sample space Ω and an associated sigma algebra $\mathcal{B}$, a probability function is a function P with domain $\mathcal{B}$ that satisfies

- **Nonnegativity:** $P(A) \geq 0$ for all $A \in \mathcal{B}$
- **Normalization:** $P(\Omega) = 1$ and $P(\emptyset) = 0$;
- **Additivity:** If $A_1, A_2, \dots \in \mathcal{B}$ are pairwise disjoint, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Note that

- The three properties given in the definition are usually referred to as the Axioms of probability (or the Kolmogorov Axioms).
- Any function P that satisfies the Axioms of probability is called a probability function.

Example of defining probability

Consider the simple experiment of tossing a fair coin, so $\Omega = \{H, T\}$, hence the reasonable probability function is

$$P(\{H\}) = P(\{T\}).$$

Example of defining probability

Consider the simple experiment of tossing a fair coin, so $\Omega = \{H, T\}$, hence the reasonable probability function is

$$P(\{H\}) = P(\{T\}).$$

- $P(\{H, T\}) = P(\{H\}) + P(\{T\}) = 1;$

Example of defining probability

Consider the simple experiment of tossing a fair coin, so $\Omega = \{H, T\}$, hence the reasonable probability function is

$$P(\{H\}) = P(\{T\}).$$

- $P(\{H, T\}) = P(\{H\}) + P(\{T\}) = 1;$
- $P(\{H\}) = P(\{T\}) = \frac{1}{2};$

Example of defining probability

Consider the simple experiment of tossing a fair coin, so $\Omega = \{H, T\}$, hence the reasonable probability function is

$$P(\{H\}) = P(\{T\}).$$

- $P(\{H, T\}) = P(\{H\}) + P(\{T\}) = 1$;
- $P(\{H\}) = P(\{T\}) = \frac{1}{2}$;
- We can also define probability over an unfair coin, e.g., $P(\{H\}) = \frac{1}{3}$ and $P(\{T\}) = \frac{2}{3}$.

Outline

Probability operators

Operators

Let Ω be the sample space, A and B be two events:

1. If $A \cap B = \emptyset$, then

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

Probability operators

Operators

Let Ω be the sample space, A and B be two events:

1. If $A \cap B = \emptyset$, then
$$P(A \cup B) = P(A) + P(B);$$
2. $$P(A^c) = P(\Omega) - P(A) = 1 - P(A);$$

Probability operators

Operators

Let Ω be the sample space, A and B be two events:

1. If $A \cap B = \emptyset$, then
$$P(A \cup B) = P(A) + P(B);$$
2. $$P(A^c) = P(\Omega) - P(A) = 1 - P(A);$$
3. If A and B are independent, then
$$P(A \cap B) = P(A) \cdot P(B);$$

Probability operators

Operators

Let Ω be the sample space, A and B be two events:

1. If $A \cap B = \emptyset$, then
$$P(A \cup B) = P(A) + P(B);$$
2. $$P(A^c) = P(\Omega) - P(A) = 1 - P(A);$$
3. If A and B are independent, then
$$P(A \cap B) = P(A) \cdot P(B);$$
4. The conditional probability of A given B, denoted by $P(A|B)$, is computed as
$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Probability properties I

If P is a probability function, A and B are any two sets in $\mathcal{B}$, then

- $P(\emptyset) = 0$, where $\emptyset$ is the empty set;

Probability properties I

If P is a probability function, A and B are any two sets in $\mathcal{B}$, then

- $P(\emptyset) = 0$, where $\emptyset$ is the empty set;
- $P(A) \leq 1$;

Probability properties I

If P is a probability function, A and B are any two sets in $\mathcal{B}$, then

- $P(\emptyset) = 0$, where $\emptyset$ is the empty set;
- $P(A) \leq 1$;
- $P(B \cap A^c) = P(B) - P(A \cap B)$;

Probability properties I

If P is a probability function, A and B are any two sets in $\mathcal{B}$, then

- $P(\emptyset) = 0$, where $\emptyset$ is the empty set;
- $P(A) \leq 1$;
- $P(B \cap A^c) = P(B) - P(A \cap B)$;
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;

Probability properties I

If P is a probability function, A and B are any two sets in $\mathcal{B}$, then

- $P(\emptyset) = 0$, where $\emptyset$ is the empty set;
- $P(A) \leq 1$;
- $P(B \cap A^c) = P(B) - P(A \cap B)$;
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;
- If $A \subset B$, then $P(A) \leq P(B)$.

Probability properties II

If P is a probability function, the collection $C_1, C_2, \cdots$ is a partition of Ω, then

- $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$;

Probability properties II

If P is a probability function, the collection $C_1, C_2, \cdots$ is a partition of Ω, then

- $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$;
- $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$ for any collection $A_1, A_2, \cdots$;

To prove the second statement, we first construct a disjoint collection $A_1^*, A_2^*, \cdots$, where

$$A_1^* = A_1, A_i^* = A_i \backslash (\bigcup_{j=1}^{i-1} A_j)$$

for $i = 2, 3, \cdots$

Running example of independence

Tossing coins

We toss a coin twice (Head = H, Tail = T), then $\Omega = \{HH, HT, TH, TT\}$.
We define three events:

1. A : the first toss is H;
2. B : the second toss is H;
3. C : the first and second toss give the same results.

Hence, we have

- $P(A) = P(B) = P(C) = \frac{1}{2}$;
- $P(A \cap B) = P(A \cap C) = P(B \cap C) = \frac{|\{HH\}|}{|\Omega|} = \frac{1}{4}$;
- $P(A \cap B \cap C) = \frac{|\{HH\}|}{|\Omega|} = \frac{1}{4}$;

Independence

Definition

- Events A_1 and A_2 are **pair-wise independent** (statistically independent) if and only if

$$P(A_1 \cap A_2) = P(A_1)P(A_2);$$

- Events $A_1, A_2, \cdots, A_n$ are **mutually independent** if

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_m}),$$

  where $i_j, j = 1, 2, \cdots, m$, are integers with $1 \leq i_1 < i_2 < \cdots < i_m \leq n$ and $m \geq 2$.

Note that mutually independent must be pair-wise independent, but pair-wise independent may not imply mutually independent (shown in previous example).

Independence

Theorem

Events A and B are pair-wise independent, then

- A and B^c are pair-wise independent;
- A^c and B are pair-wise independent;
- A^c and B^c are pair-wise independent;

Proof.

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

Hence, we have A and B^c are independent. □

Conditional probability

Definition

Let A and B be events with $P(B) > 0$. The **conditional probability** of A given B, denoted by $P(A|B)$, is defined as

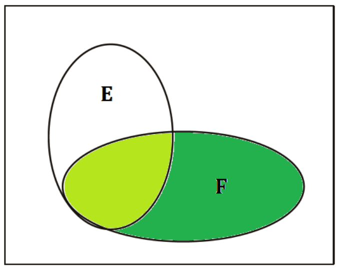
$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Conditional probability

Definition

Let A and B be events with $P(B) > 0$. The **conditional probability** of A given B, denoted by $P(A|B)$, is defined as

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

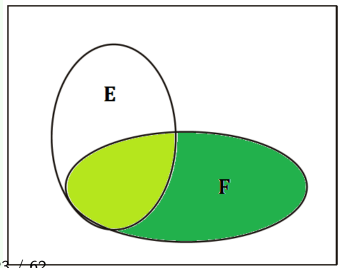


Conditional probability

Definition

Let A and B be events with $P(B) > 0$. The **conditional probability** of A given B, denoted by $P(A|B)$, is defined as

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



- $P(A|B)$ is the probability of A, given that B occurred
- B is our new sample space;
- $P(A|B)$ is undefined if $P(B) = 0$.

Example

Question: What is the conditional probability that a family with two children has two boys, given they have at least one boy? Assume that each of the possibilities BB, BG, GB, and GG is equally likely.

Example

Question: What is the conditional probability that a family with two children has two boys, given they have at least one boy? Assume that each of the possibilities BB, BG, GB, and GG is equally likely.

Solution: Let A be the event that a family with two children has two boys, and let B be the event that a family with two children has at least one boy.

Example

Question: What is the conditional probability that a family with two children has two boys, given they have at least one boy? Assume that each of the possibilities BB, BG, GB, and GG is equally likely.

Solution: Let A be the event that a family with two children has two boys, and let B be the event that a family with two children has at least one boy.
Thus, $A = \{BB\}$, $B = \{BB, BG, GB\}$, and $A \cap B = \{BB\}$.

Example

Question: What is the conditional probability that a family with two children has two boys, given they have at least one boy? Assume that each of the possibilities BB, BG, GB, and GG is equally likely.

Solution: Let A be the event that a family with two children has two boys, and let B be the event that a family with two children has at least one boy.
Thus, $A = \{BB\}$, $B = \{BB, BG, GB\}$, and $A \cap B = \{BB\}$.
Because the four possibilities are equally likely, it follows that $P(B) = 3/4$ and $P(A \cap B) = 1/4$.

Example

Question: What is the conditional probability that a family with two children has two boys, given they have at least one boy? Assume that each of the possibilities BB, BG, GB, and GG is equally likely.

Solution: Let A be the event that a family with two children has two boys, and let B be the event that a family with two children has at least one boy.
Thus, $A = \{BB\}$, $B = \{BB, BG, GB\}$, and $A \cap B = \{BB\}$.
Because the four possibilities are equally likely, it follows that $P(B) = 3/4$ and $P(A \cap B) = 1/4$.
We conclude that

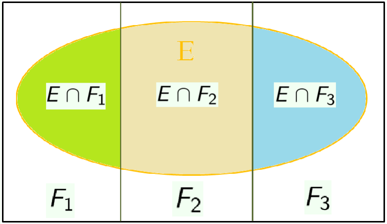
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Remarks for conditional probability

- $P(A|B) = P(A)$ if events A and B are independent;

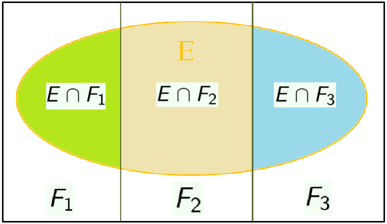
Remarks for conditional probability

- $P(A|B) = P(A)$ if events A and B are independent;
- $P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$;



Remarks for conditional probability

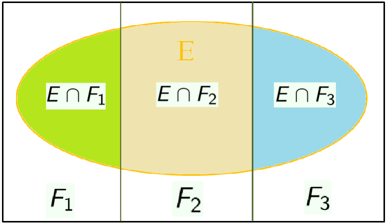
- $P(A|B) = P(A)$ if events A and B are independent;
- $P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$;



- $P(A) = P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2) + P(B_3) \cdot P(A|B_3)$ if B_1, B_2 and B_3 is a partition of Ω (**Total probability theorem**);

Remarks for conditional probability

- $P(A|B) = P(A)$ if events A and B are independent;
- $P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$;



- $P(A) = P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2) + P(B_3) \cdot P(A|B_3)$ if B_1, B_2 and B_3 is a partition of Ω (**Total probability theorem**);
- $P(B_i|A) = \frac{P(B_i) \cdot P(A|B_i)}{P(A)} = \frac{P(B_i) \cdot P(A|B_i)}{\sum_i P(B_i) \cdot P(A|B_i)}$ (**Bayes rule**).

Proof of the total probability theorem

Theorem

Let B_i (for $i = 1, 2, \cdots, n$) be a partition of sample space Ω, for any event A, then

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

Proof of the total probability theorem

Theorem

Let B_i (for $i = 1, 2, \cdots, n$) be a partition of sample space Ω, for any event A, then

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

Proof:

$$\begin{aligned} P(A) &= P(A \cap \Omega) = P(A \cap (B_1 \cup B_2 \cup \cdots \cup B_n)) \\ &= P((A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_n)) \\ &= P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n) \\ &= \sum_{i=1}^{n} P(B_i) \cdot P(A|B_i). \end{aligned}$$

Bayes' Theorem

Theorem

Suppose that A is an event from a sample space Ω and $B_1, B_2, \cdots, B_n$ is a partition of the sample space. Let $P(A) \neq 0$ and $P(B_i) \neq 0$ for $\forall i$. Then

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{k=1}^{n} P(A|B_k)P(B_k)}.$$

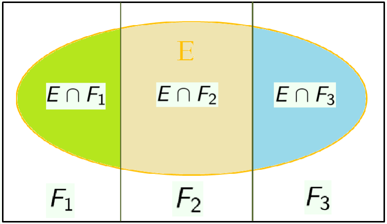
Bayes' Theorem

Theorem

Suppose that A is an event from a sample space Ω and $B_1, B_2, \cdots, B_n$ is a partition of the sample space. Let $P(A) \neq 0$ and $P(B_i) \neq 0$ for $\forall i$. Then

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{k=1}^{n} P(A|B_k)P(B_k)}.$$

Proof:



Diagnostic test for rare disease

Suppose that one of 100,000 persons has a particular rare disease for which there is a fairly accurate diagnostic test. This test is correct 99.0% when given to a person selected at random who has the disease; it is correct 99.5% when given to a person selected at random who does not have the disease. Please find

- the probability that a person who tests positive for the disease has the disease?
- the probability that a person who tests negative for the disease does not have the disease?

Should a person who tests positive be very concerned that he or she has the disease?

Diagnostic test for rare disease

Suppose that one of 100,000 persons has a particular rare disease for which there is a fairly accurate diagnostic test. This test is correct 99.0% when given to a person selected at random who has the disease; it is correct 99.5% when given to a person selected at random who does not have the disease. Please find

- the probability that a person who tests positive for the disease has the disease?
- the probability that a person who tests negative for the disease does not have the disease?

Should a person who tests positive be very concerned that he or she has the disease?

Solution: Let B be the event that a person selected at random has the disease, and let A be the event that a person selected at random tests positive for the disease.

Diagnostic test for rare disease Cont'd

Hence, we have $p(B) = 1/100,000 = 10^{-5}$. Then we also have $P(A|B) = 0.99$, $P(A^c|B) = 0.01$, $P(A^c|B^c) = 0.995$, and $P(A|B^c) = 0.005$.

Diagnostic test for rare disease Cont'd

Hence, we have $p(B) = 1/100,000 = 10^{-5}$. Then we also have $P(A|B) = 0.99$, $P(A^c|B) = 0.01$, $P(A^c|B^c) = 0.995$, and $P(A|B^c) = 0.005$.

Case a: In terms of Bayes' theorem, we have

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$
$$= \frac{0.99 \cdot 10^{-5}}{0.99 \cdot 10^{-5} + 0.005 \cdot 0.99999} \approx 0.002$$

Diagnostic test for rare disease Cont'd

Hence, we have $p(B) = 1/100,000 = 10^{-5}$. Then we also have $P(A|B) = 0.99$, $P(A^c|B) = 0.01$, $P(A^c|B^c) = 0.995$, and $P(A|B^c) = 0.005$.

Case a: In terms of Bayes' theorem, we have

$$\begin{aligned} P(B|A) &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \\ &= \frac{0.99 \cdot 10^{-5}}{0.99 \cdot 10^{-5} + 0.005 \cdot 0.99999} \approx 0.002 \end{aligned}$$

Diagnostic test for rare disease Cont'd

Hence, we have $p(B) = 1/100,000 = 10^{-5}$. Then we also have $P(A|B) = 0.99$, $P(A^c|B) = 0.01$, $P(A^c|B^c) = 0.995$, and $P(A|B^c) = 0.005$.

Case a: In terms of Bayes' theorem, we have

$$\begin{aligned} P(B|A) &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \\ &= \frac{0.99 \cdot 10^{-5}}{0.99 \cdot 10^{-5} + 0.005 \cdot 0.99999} \approx 0.002 \end{aligned}$$

Case b: Similarly, we have

$$P(B^c|A^c) = \frac{P(A^c|B^c)P(B^c)}{P(A^c|B^c)P(B^c) + P(A^c|F)P(F)} \approx 0.9999999$$

Bayesian spam filters

Most electronic mailboxes receive a flood of unwanted and unsolicited messages, known as **spam**.

Bayesian spam filters

Most electronic mailboxes receive a flood of unwanted and unsolicited messages, known as **spam**.

On the Internet, an **Internet Water Army** is a group of Internet ghostwriters paid to post online comments with particular content.

Bayesian spam filters

Most electronic mailboxes receive a flood of unwanted and unsolicited messages, known as **spam**.

On the Internet, an **Internet Water Army** is a group of Internet ghostwriters paid to post online comments with particular content.

Question: How to detect spam email?

Bayesian spam filters

Most electronic mailboxes receive a flood of unwanted and unsolicited messages, known as **spam**.

On the Internet, an **Internet Water Army** is a group of Internet ghostwriters paid to post online comments with particular content.

Question: How to detect spam email?

Solution: Bayesian spam filters look for occurrences of particular words in messages. For a particular word w, the probability that w appears in a spam e-mail message is estimated by determining $\#$ times w appears in a message from a large set of messages known to be spam and $\#$ times it appears in a large set of messages known not to be spam.

Bayesian spam filters

Most electronic mailboxes receive a flood of unwanted and unsolicited messages, known as **spam**.
On the Internet, an **Internet Water Army** is a group of Internet ghostwriters paid to post online comments with particular content.

Question: How to detect spam email?
Solution: Bayesian spam filters look for occurrences of particular words in messages. For a particular word w, the probability that w appears in a spam e-mail message is estimated by determining $\#$ times w appears in a message from a large set of messages known to be spam and $\#$ times it appears in a large set of messages known not to be spam.
Step 1: Collect ground-truth Suppose we have a set B of messages known to be spam and a set G of messages known not to be spam.

Bayesian spam filters Cont'd

Step 2: Learn parameters We next identify the words that occur in B and in G. Let $n_B(w)$ and $n_G(w)$ be # messages containing word w in sets B and G, respectively.

Bayesian spam filters Cont'd

Step 2: Learn parameters We next identify the words that occur in B and in G. Let $n_B(w)$ and $n_G(w)$ be # messages containing word w in sets B and G, respectively.

Let $p(w) = n_B(w)/|B|$ and $q(w) = n_G(w)/|G|$ be the empirical probabilities that a message are not spam and spam contains word w, respectively.

Step 3: Make decision Now suppose we receive a new e-mail message containing word w. Let B be the event that the message is spam. Let A be the event that the message contains word w.

Bayesian spam filters Cont'd

Step 2: Learn parameters We next identify the words that occur in B and in G. Let $n_B(w)$ and $n_G(w)$ be # messages containing word w in sets B and G, respectively.

Let $p(w) = n_B(w)/|B|$ and $q(w) = n_G(w)/|G|$ be the empirical probabilities that a message are not spam and spam contains word w, respectively.

Step 3: Make decision Now suppose we receive a new e-mail message containing word w. Let B be the event that the message is spam. Let A be the event that the message contains word w.

By Bayes theorem, the probability that the message is spam, given that it contains word w, is

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

Bayesian spam filters Cont'd

To apply the above formula, we first estimate $P(B)$ (probability of spam) and $P(B^c)$ (probability of not spam).

Bayesian spam filters Cont'd

To apply the above formula, we first estimate $P(B)$ (probability of spam) and $P(B^c)$ (probability of not spam).

Without prior knowledge, for simplicity we assume that the message is equally likely to be spam as it is not to be spam, i.e., $P(B) = P(B^c) = 1/2$.

Bayesian spam filters Cont'd

To apply the above formula, we first estimate $P(B)$ (probability of spam) and $P(B^c)$ (probability of not spam).

Without prior knowledge, for simplicity we assume that the message is equally likely to be spam as it is not to be spam, i.e., $P(B) = P(B^c) = 1/2$.

Using this assumption, we find that the probability that a message is spam, given that it contains word w, is

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

Bayesian spam filters Cont'd

To apply the above formula, we first estimate $P(B)$ (probability of spam) and $P(B^c)$ (probability of not spam).

Without prior knowledge, for simplicity we assume that the message is equally likely to be spam as it is not to be spam, i.e., $P(B) = P(B^c) = 1/2$.

Using this assumption, we find that the probability that a message is spam, given that it contains word w, is

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

$P(A|B)$ and $P(A|B^c)$ are known, $P(B|A)$ can be estimated by

$$r(w) = \frac{p(w)}{p(w) + q(w)}.$$

Extended Bayesian spam filters

The more words we use to estimate the probability that an incoming mail message is spam, the better is our chance that we correctly determine whether it is spam.

In general, if A_i is the event that the message contains word w_i , assuming that $P(S) = P(S^c)$, and that events $A_i|S$ are independent, then by Bayes theorem the probability that a message containing all words $w_1, w_2, \cdots, w_k$ is spam is

$$
\begin{aligned}
P(S|\bigcap_{i=1}^{k} A_i) &= \frac{P(\bigcap_{i=1}^{k} A_i|S)P(S)}{P(\bigcap_{i=1}^{k} A_i|S)P(S) + P(\bigcap_{i=1}^{k} A_i|\overline{S})P(\overline{S})} \\
&= \frac{\Pi_{i=1}^{k} P(A_i|S)}{\Pi_{i=1}^{k} P(A_i|S) + \Pi_{i=1}^{k} P(A_i|\overline{S})} \\
&\approx \frac{\Pi_{i=1}^{k} p(w_i)}{\Pi_{i=1}^{k} p(w_i) + \Pi_{i=1}^{k} q(w_i)} = r(w_1, w_2, \cdots, w_k).
\end{aligned}
$$

Naive Bayes

- Why is this called Naive Bayes?

Naive Bayes

- Why is this called Naive Bayes?
- The model employs the chain rule for repeated applications of the definition of conditional probability.

Naive Bayes

- Why is this called Naive Bayes?
- The model employs the chain rule for repeated applications of the definition of conditional probability.
- To handle underflow, we calculate $\Pi_{i=1}^{n} P(X_i|S) = exp(\sum_{i=1}^{n} \log P(X_i|S))$.

Outline

Counting principle I

Product rule

Suppose that a procedure consists of a sequence of two tasks. If there are n_1 ways to do the first task, there are n_2 ways to do the second task, then there are $n_1 n_2$ ways to do the procedure.

Counting principle I

Product rule

Suppose that a procedure consists of a sequence of two tasks. If there are n_1 ways to do the first task, there are n_2 ways to do the second task, then there are $n_1 n_2$ ways to do the procedure.

Extended version: A procedure is followed by tasks $T_1, T_2, \cdots, T_m$ in sequence. If each task T_i can be done in n_i ways independently, then there are $n_1 \cdot n_2 \cdot \cdots \cdot n_m$ ways to carry out the procedure.

Counting principle I

Product rule

Suppose that a procedure consists of a sequence of two tasks. If there are n_1 ways to do the first task, there are n_2 ways to do the second task, then there are $n_1 n_2$ ways to do the procedure.

Extended version: A procedure is followed by tasks $T_1, T_2, \cdots, T_m$ in sequence. If each task T_i can be done in n_i ways independently, then there are $n_1 \cdot n_2 \cdot \cdots \cdot n_m$ ways to carry out the procedure.

Sum rule

If a task can be done either in one of n_1 ways or in one of n_2 ways, where none of the set of n_1 ways is the same as any of the set of n_2 ways, then there are $n_1 + n_2$ ways to do the task.

Counting principle I

Product rule

Suppose that a procedure consists of a sequence of two tasks. If there are n_1 ways to do the first task, there are n_2 ways to do the second task, then there are $n_1 n_2$ ways to do the procedure.

Extended version: A procedure is followed by tasks $T_1, T_2, \cdots, T_m$ in sequence. If each task T_i can be done in n_i ways independently, then there are $n_1 \cdot n_2 \cdot \cdots \cdot n_m$ ways to carry out the procedure.

Sum rule

If a task can be done either in one of n_1 ways or in one of n_2 ways, where none of the set of n_1 ways is the same as any of the set of n_2 ways, then there are $n_1 + n_2$ ways to do the task.

Extended version: A procedure can be done by m ways, each way W_i has n_i possibilities, then there are $\sum_{i=1}^{m} n_i$ ways for the procedure.

Counting principle II

Subtraction rule

There are n/d ways to do a task if it can be done using a procedure that can be carried out in n ways, and for every way w, exactly d of the n ways correspond to way w.

Counting principle II

Subtraction rule

There are n/d ways to do a task if it can be done using a procedure that can be carried out in n ways, and for every way w, exactly d of the n ways correspond to way w.

Subtraction rule

If a task can be done in either n_1 ways or n_2 ways, then the number of ways to do the task is $n_1 + n_2$ minus the number of ways to do the task that are common to the two different ways. The rule is also called the principle of **inclusion-exclusion**, i.e.,

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

Permutations and combinations

Permutations

An ordered arrangement of m elements of a set is called an m-permutation. # m-permutations of a set with n distinct elements is

$$P(n,m) = n(n-1)(n-2)\cdot\cdots\cdot(n-m+1) = \frac{n!}{(n-m)!}.$$

Permutations and combinations

Permutations

An ordered arrangement of m elements of a set is called an m-permutation. $\#$ m-permutations of a set with n distinct elements is

$$P(n, m) = n(n-1)(n-2) \cdot \dots \cdot (n-m+1) = \frac{n!}{(n-m)!}.$$

Combinations

An m-combination of elements of a set is an unordered selection of m elements from the set. $\#$ m-combinations of a set with n elements equals

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

where $C(n, m)$ is also denoted as $\binom{n}{m}$.

Finite probability

If S is a finite nonempty sample space of equally likely outcomes, and E is an event, that is, a subset of S, then the probability of E is

$$p(E) = \frac{|E|}{|S|}.$$

Finite probability

If S is a finite nonempty sample space of equally likely outcomes, and E is an event, that is, a subset of S, then the probability of E is

$$p(E) = \frac{|E|}{|S|}.$$

- Let all outcomes be equally likely;
- Computing probabilities $\equiv$ two countings;
  - Counting the successful ways of the event;
  - Counting the size of the sample space;

Examples

Example I

Question: An urn contains four blue balls and five red balls. What is the probability that a ball chosen at random from the urn is blue?

Solution: Let S be the sample space, i.e.,

$$S = \{{\color{blue}\bigcirc}_1, {\color{blue}\bigcirc}_2, {\color{blue}\bigcirc}_3, {\color{blue}\bigcirc}_4, {\color{red}\bigcirc}_1, {\color{red}\bigcirc}_2, {\color{red}\bigcirc}_3, {\color{red}\bigcirc}_4, {\color{red}\bigcirc}_5\}.$$

Let E be the event of choosing a blue ball, i.e.,

$$E = \{{\color{blue}\bigcirc}_1, {\color{blue}\bigcirc}_2, {\color{blue}\bigcirc}_3, {\color{blue}\bigcirc}_4\}.$$

In terms of the definition, we can compute the probability as

$$P(E) = \frac{|E|}{|S|} = \frac{4}{9}.$$

Examples Cont'd

Example II

Question: What is the probability that when two dice are rolled, the sum of the numbers on the two dice is 7?

Solution:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Examples Cont'd

Example II

Question: What is the probability that when two dice are rolled, the sum of the numbers on the two dice is 7?

Solution:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

- There are a total of 36 possible outcomes when two dice are rolled.

Examples Cont'd

Example II

Question: What is the probability that when two dice are rolled, the sum of the numbers on the two dice is 7?

Solution:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

- There are a total of 36 possible outcomes when two dice are rolled.
- There are six successful outcomes, namely, $(1,6), (2,5)$, $(3,4), (4,3)$, $(5,2)$, and $(6,1)$.
- Hence, the probability that a seven comes up when two fair dice are rolled is $6/36 = 1/6$.

Examples Cont'd

Example III

Question: In a lottery, players win a large prize when they pick four random digits that match, in the correct order. A smaller prize is won if only three digits are matched. What is the probability that a player wins the large prize? What is the probability that a player wins the small prize?

Examples Cont'd

Example III

Question: In a lottery, players win a large prize when they pick four random digits that match, in the correct order. A smaller prize is won if only three digits are matched. What is the probability that a player wins the large prize? What is the probability that a player wins the small prize?

Solution: By the product rule, there are $10^4 = 10,000$ ways to choose four digits.

Examples Cont'd

Example III

Question: In a lottery, players win a large prize when they pick four random digits that match, in the correct order. A smaller prize is won if only three digits are matched. What is the probability that a player wins the large prize? What is the probability that a player wins the small prize?

Solution: By the product rule, there are $10^4 = 10,000$ ways to choose four digits.

Large prize case: There is only one way to choose all four digits correctly. Thus, the probability is $1/10,000 = 0.0001$.

Examples Cont'd

Example III

Question: In a lottery, players win a large prize when they pick four random digits that match, in the correct order. A smaller prize is won if only three digits are matched. What is the probability that a player wins the large prize? What is the probability that a player wins the small prize?

Solution: By the product rule, there are $10^4 = 10,000$ ways to choose four digits.

Large prize case: There is only one way to choose all four digits correctly. Thus, the probability is $1/10,000 = 0.0001$.

Small prize case: Exactly one digit must be wrong to get three digits correct, but not all four correct.

Examples Cont'd

Example III

Question: In a lottery, players win a large prize when they pick four random digits that match, in the correct order. A smaller prize is won if only three digits are matched. What is the probability that a player wins the large prize? What is the probability that a player wins the small prize?

Solution: By the product rule, there are $10^4 = 10,000$ ways to choose four digits.

Large prize case: There is only one way to choose all four digits correctly. Thus, the probability is $1/10,000 = 0.0001$.

Small prize case: Exactly one digit must be wrong to get three digits correct, but not all four correct. Hence, there is a total of $\binom{4}{1} \times 9 = 36$ ways to choose four digits with exactly three of the four digits correct. Thus, the probability that a player wins the smaller prize is $36/10,000 = 9/2500 = 0.0036$.

Examples Cont'd

Example IV

Question: Find the probabilities that a poker hand contains four cards of one kind, or a full house (i.e., three of one kind and two of another kind).

Examples Cont'd

Example IV

Question: Find the probabilities that a poker hand contains four cards of one kind, or a full house (i.e., three of one kind and two of another kind).

Solution: There are $C(52, 5)$ different hands of five cards.

Examples Cont'd

Example IV

Question: Find the probabilities that a poker hand contains four cards of one kind, or a full house (i.e., three of one kind and two of another kind).

Solution: There are $C(52, 5)$ different hands of five cards.

Case I: # hands of five cards with four cards of one kind is

$$C(13, 1)C(4, 4)C(48, 1).$$

Examples Cont'd

Example IV

Question: Find the probabilities that a poker hand contains four cards of one kind, or a full house (i.e., three of one kind and two of another kind).

Solution: There are $C(52,5)$ different hands of five cards.

Case I: # hands of five cards with four cards of one kind is

$$C(13,1)C(4,4)C(48,1).$$

Case II: # hands of three of one kind and two of another kind is

$$P(13,2)C(4,3)C(4,2).$$

Examples Cont'd

Example IV

Question: Find the probabilities that a poker hand contains four cards of one kind, or a full house (i.e., three of one kind and two of another kind).

Solution: There are $C(52,5)$ different hands of five cards.

Case I: # hands of five cards with four cards of one kind is

$$C(13,1)C(4,4)C(48,1).$$

Case II: # hands of three of one kind and two of another kind is

$$P(13,2)C(4,3)C(4,2).$$

Hence, the probabilities are

$$\frac{C(13,1)C(4,4)C(48,1)}{C(52,5)} \approx 0.00024, \frac{C(13,2)C(4,3)C(4,2)}{C(52,5)} \approx 0.0014.$$

Outline

Random variables

Definition: A **random variable** (r.v.) X is a function from sample space Ω of an experiment to the set of real numbers in R, i.e.,

$$\forall \omega \in \Omega, X(\omega) = x \in R.$$

Random variables

Definition: A **random variable** (r.v.) X is a function from sample space Ω of an experiment to the set of real numbers in R, i.e.,

$$\forall \omega \in \Omega, X(\omega) = x \in R.$$

Remarks

- Note that a random variable is a function. It is not a variable, and it is not random!
- We usually use notation X, Y, etc. to represent a r.v., and x, y to represent the numerical values. For example, $X = x$ means that r.v. X has value x.
- The domain of the function can be countable and uncountable. If it is countable, the random variable is a discrete r.v., otherwise continuous r.v..

Examples of r.v.

A coin is tossed. If X is the r.v. whose value is the number of heads obtained, then $X(H) = 1, X(T) = 0$.

Examples of r.v.

A coin is tossed. If X is the r.v. whose value is the number of heads obtained, then $X(H) = 1, X(T) = 0$.

And then tossed again. We define sample space $\Omega = \{HH, HT, TH, TT\}$. If Y is the r.v. whose value is the number of heads obtained, then

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

Examples of r.v.

A coin is tossed. If X is the r.v. whose value is the number of heads obtained, then $X(H) = 1, X(T) = 0$.

And then tossed again. We define sample space $\Omega = \{HH, HT, TH, TT\}$. If Y is the r.v. whose value is the number of heads obtained, then

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

When a player rolls a die, he will win \$1 if the outcome is 1,2 or 3, otherwise lose 1\$. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ and define X as follows:

$$X(1) = X(2) = X(3) = 1, X(4) = X(5) = X(6) = -1.$$

Random variables VS. events

Suppose now that a sample space $\Omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$ is given, and r.v. X on Ω is defined the number of heads obtained when we toss a coin twice.

Random variables VS. events

Suppose now that a sample space $\Omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$ is given, and r.v. X on Ω is defined the number of heads obtained when we toss a coin twice.

- Event E_1 represents only one head obtained. Hence,

$$E_1 = \{\omega : X(\omega) = 1\};$$

- Event E_2 represents even heads obtained. Hence,

$$E = \{\omega : X(\omega) \mod 2 = 0\};$$

- Event E_2 represents at least one heads obtained. Hence,

$$E = \{\omega : X(\omega) > 0\}.$$

These indicate that we can also define probability about r.v.s.

Outline

Cumulative distribution function

The **cumulative distribution function** or cdf of a r.v. X, denoted by $F_X(x)$ is defined by

$$F_X(x) = P_X(X \leq x),$$

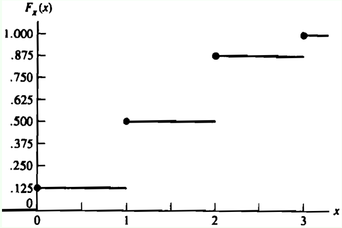
for all x.

Cumulative distribution function

The **cumulative distribution function** or cdf of a r.v. X, denoted by $F_X(x)$ is defined by

$$F_X(x) = P_X(X \leq x),$$

for all x.

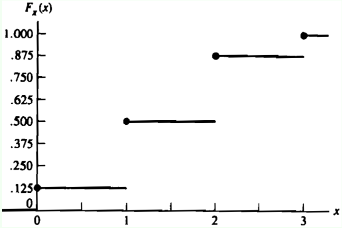


Cumulative distribution function

The **cumulative distribution function** or cdf of a r.v. X, denoted by $F_X(x)$ is defined by

$$F_X(x) = P_X(X \leq x),$$

for all x.



Consider the experiment of tossing three fair coins, and let $X = \#$ heads observed. The cdf of X is

$$F_X(x) = \begin{cases} 0, & \text{if } -\infty < x < 0; \\ \frac{1}{8}, & \text{if } 0 \leq x < 1; \\ \frac{1}{2}, & \text{if } 1 \leq x < 2; \\ \frac{7}{8}, & \text{if } 2 \leq x < 3; \\ 1, & \text{if } 3 \leq x < +\infty. \end{cases}$$

Theorem

The function $F(x)$ is a cdf if and only if the following three conditions hold:

- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$;
- $F(x)$ is a nondecreasing function of x;
- $F(x)$ is right-continuous; that is, for ever number x_0, $\lim_{x \downarrow x_0} F(x) = F(x_0)$.

Theorem

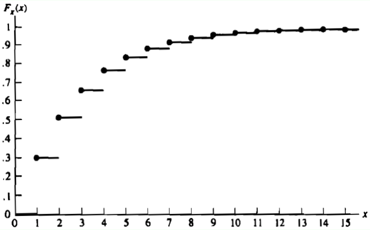
The function $F(x)$ is a cdf if and only if the following three conditions hold:

- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$;
- $F(x)$ is a nondecreasing function of x;
- $F(x)$ is right-continuous; that is, for ever number x_0, $\lim_{x \downarrow x_0} F(x) = F(x_0)$.

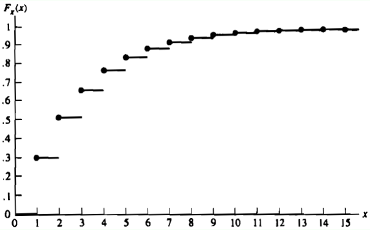
Example

Suppose we do an experiment that consists of tossing a coin until a head appears. Let $p =$ probability of a head on any given toss, and define a r.v. $X = \#$ tosses required to get a head. Then for any $x = 1, 2, \cdot$, $P(X \leq x) = \sum_{i=1}^{x} (1-p)^{i-1} p = 1 - (1-p)^x$.

Example Cont'd



Example Cont'd



Given $0 < p < 1$, we have

- $\lim_{x\to-\infty} 1-(1-p)^x = 0$ and $\lim_{x\to+\infty} 1-(1-p)^x = 1$;
- $1-(1-p)^x$ is a nondecreasing function of x;
- For any x, $F(x)(x+\epsilon) = F_X(x)$ if $\epsilon > 0$ is sufficiently small. Hence,

$$\lim_{\epsilon\downarrow 0} F_X(x+\epsilon) = F_X(x).$$

$F_X(x)$ is the cdf of a distribution called the **Geometric distribution**.

Example Cont'd

An example of a continuous cdf is the function

$$F_X(x) = \frac{1}{1+e^{-x}}.$$

- $\lim_{x\to-\infty} F_X(x) = 0$ since $\lim_{x\to-\infty} e^{-x} = \infty$;
- $\lim_{x\to+\infty} F_X(x) = 1$ since $\lim_{x\to+\infty} e^{-x} = 0$;
- Differentiating $F_X(x)$ gives

$$\frac{d}{dx} F_X(x) = \frac{e^{-x}}{(1+e^{-x})^2} > 0;$$

- $F_X(x)$ is not only right-continuous, but also continuous.

This is a special case of the **logistic distribution**.

Continuous and discrete r.v.s

A r.v. X is continuous if $F_X(x)$ is a continuous function of x.
And a r.v. X is discrete if $F_X(x)$ is a step function of x.

Continuous and discrete r.v.s

A r.v. X is continuous if $F_X(x)$ is a continuous function of x. And a r.v. X is discrete if $F_X(x)$ is a step function of x.

The r.v.s X and Y are identically distributed if, for every set $A \in \mathcal{B}$

$$P(X(\omega \in A)) = P(Y(\omega \in A)).$$

Note that two r.v.s that are identically distributed are not necessarily equal.

Continuous and discrete r.v.s

A r.v. X is continuous if $F_X(x)$ is a continuous function of x. And a r.v. X is discrete if $F_X(x)$ is a step function of x.

The r.v.s X and Y are identically distributed if, for every set $A \in \mathcal{B}$

$$P(X(\omega \in A)) = P(Y(\omega \in A)).$$

Note that two r.v.s that are identically distributed are not necessarily equal.

Theorem

The following two statements are equivalent:

- The r.v., X and Y are identically distributed;
- $F_X(x) = F_Y(x)$ for all x.

Example

Note that two r.v.s that are identically distributed are not necessarily equal.

Example

Note that two r.v.s that are identically distributed are not necessarily equal.

For example, consider an example of tossing a fair coin three times. Define r.v.s

$$X = \# \text{ heads observed}$$

and

$$Y = \# \text{ tails observed}.$$

We have $P(X = k) = P(Y = k)$, i.e., X and Y are identically distributed. However, we do not have $X(\omega) = Y(\omega)$ for any $\omega \in \Omega$.

Outline

Probability mass function

For all x, the **probability mass function** (pmf) of a discrete r.v. X on Ω is given by $f_X(x) = P(X = x)$.

Probability mass function

For all x, the **probability mass function** (pmf) of a discrete r.v. X on Ω is given by $f_X(x) = P(X = x)$.

For the geometric distribution, we have the pmf

$$f_X(x) = P(X = x) = \begin{cases} (1-p)^{x-1}p, & \text{for } x = 1, 2, \cdots \\ 0, & \text{otherwise.} \end{cases}$$

- Recall that $P(X = x)$ or, equivalently, $f_X(x)$ is the size of the jump in the cdf at x.
- We can use the pmf to calculate probabilities, for positive integers a and $b \geq a$, we have

$$P(a \leq X \leq b) = \sum_{k=a}^{b} f_X(k) = \sum_{k=a}^{b} (1-p)^{k-1}p.$$

Probability mass function

For all x, the **probability mass function** (pmf) of a discrete r.v. X on Ω is given by $f_X(x) = P(X = x)$.

Probability mass function

For all x, the **probability mass function** (pmf) of a discrete r.v. X on Ω is given by $f_X(x) = P(X = x)$.

For the geometric distribution, we have the pmf

$$f_X(x) = P(X = x) = \begin{cases} (1-p)^{x-1}p, & \text{for } x = 1, 2, \cdots \\ 0, & \text{otherwise.} \end{cases}$$

- Recall that $P(X = x)$ or, equivalently, $f_X(x)$ is the size of the jump in the cdf at x.
- We can use the pmf to calculate probabilities, for positive integers a and $b \geq a$, we have

$$P(a \leq X \leq b) = \sum_{k=a}^{b} f_X(k) = \sum_{k=a}^{b} (1-p)^{k-1} p.$$

Examples of pmf

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$		

Examples of pmf

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$

Examples of pmf

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

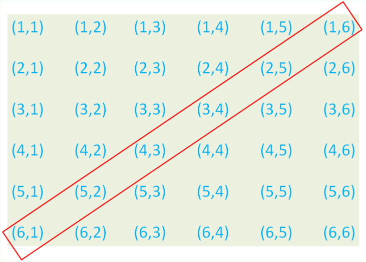
(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$		

Examples of pmf

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

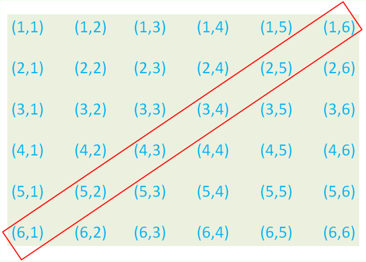


Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$

Examples of pmf

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

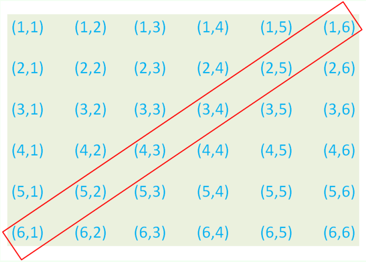


Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$		

Examples of pmf

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?



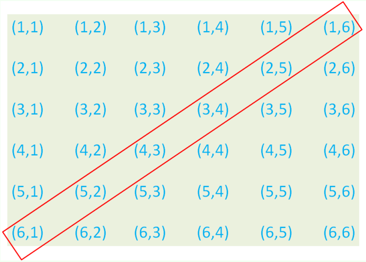
(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$	7	$\frac{1}{6}$

Examples of pmf

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?



Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$	7	$\frac{1}{6}$
8	$\frac{5}{36}$		

Examples of pmf

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

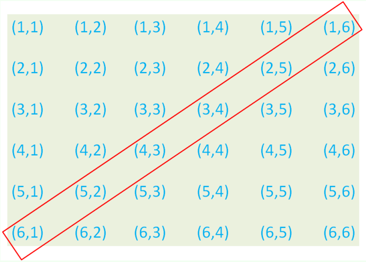
(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$	7	$\frac{1}{6}$
8	$\frac{5}{36}$	9	$\frac{1}{9}$

Examples of pmf

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

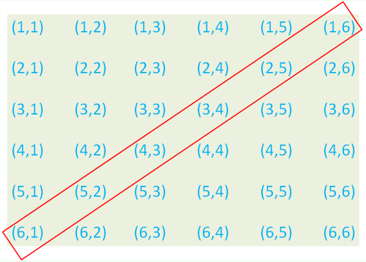


Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$	7	$\frac{1}{6}$
8	$\frac{5}{36}$	9	$\frac{1}{9}$
10	$\frac{1}{12}$		

Examples of pmf

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?



Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$	7	$\frac{1}{6}$
8	$\frac{5}{36}$	9	$\frac{1}{9}$
10	$\frac{1}{12}$	11	$\frac{1}{18}$

Examples of pmf

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$	7	$\frac{1}{6}$
8	$\frac{5}{36}$	9	$\frac{1}{9}$
10	$\frac{1}{12}$	11	$\frac{1}{18}$
12	$\frac{1}{36}$		

Probability density function

If we naively try to calculate $P(X = x)$ for a continuous r.v. and any $\epsilon > 0$, we have

$$P(X = x) \leq P(x - \epsilon < X \leq x) = F_X(x) - F_X(x - \epsilon).$$

Therefore,

$$0 \leq P(X = x) \leq \lim_{\epsilon \downarrow 0} [F_X(x) - F_X(x - \epsilon)] = 0.$$

Probability density function

If we naively try to calculate $P(X = x)$ for a continuous r.v. and any $\epsilon > 0$, we have

$$P(X = x) \leq P(x - \epsilon < X \leq x) = F_X(x) - F_X(x - \epsilon).$$

Therefore,

$$0 \leq P(X = x) \leq \lim_{\epsilon \downarrow 0}[F_X(x) - F_X(x - \epsilon)] = 0.$$

Definition

The probability density function or pdf, $f_X(x)$, of a continuous r.v. X is the function that satisfies

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

Remarks of pdf

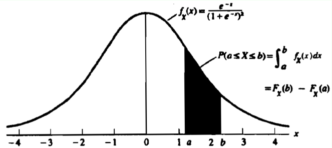
- "X has a distribution given by $F_X(x)$" is abbreviated symbolically by "$X \sim F_X(x)$" ($X \sim f_X(x)$).
- Given a continuous r.v. X, since $P(X = x) = 0$, we have

$$P(a < X < b) = P(a \le X < b) = P(a \le X \le b) = P(a < X \le b).$$

Remarks of pdf

- "X has a distribution given by $F_X(x)$" is abbreviated symbolically by "$X \sim F_X(x)$" ($X \sim f_X(x)$).
- Given a continuous r.v. X, since $P(X = x) = 0$, we have

$$P(a < X < b) = P(a \le X < b) = P(a \le X \le b) = P(a < X \le b).$$



For logistic distribution, $F_x(x) = \frac{1}{1+e^{-x}}$.
We have

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{e^{-x}}{(1+e^{-x})^2}.$$

$$P(a < X < b) = F_X(b) - F_X(a) = \int_{-\infty}^{b} f_X(t)dt - \int_{-\infty}^{a} f_X(t)dt$$

$$= \int_a^b f_X(t)dt.$$

Theorem

A function $f_X(x)$ is a pdf (or pmf) of a r.v. X if and only if

- $f_X(x) \geq 0$ for all x;
- $\sum_x f_X(x) = 1($ pmf$)$ or $\int_{-\infty}^{+\infty} f_X(x)dx = 1($ pdf$)$.

Theorem

A function $f_X(x)$ is a pdf (or pmf) of a r.v. X if and only if

- $f_X(x) \geq 0$ for all x;
- $\sum_x f_X(x) = 1($ pmf$)$ or $\int_{-\infty}^{+\infty} f_X(x)dx = 1($ pdf$)$.

Actually, any nonnegative function with a finite positive integral (or sum) can be turned into a pdf or pmf.
For example, if $h(x)$ is any nonnegative function that is positive on a set A, 0 otherwise, and

$$\int_{\{x \in A\}} h(x)dx = K < \infty$$

for some constant $K > 0$, then the function $f_X(x) = \frac{h(x)}{K}$ is a pdf of a r.v. X taking values in A.

Take-aways

Conclusions

- Introduction
- Set Theory
- Basics of Probability Theory
  - The Calculus of Probabilities
  - Counting
- Random Variable and Distributions
  - Random Variable
  - Distribution Functions
  - Density and Mass Functions