

Probability and Statistics

Omid Safarzadeh

January 8, 2022

Table of contents

- 1 Sample Space
- 2 Probability Theory Foundation
 - Axiomatic Foundations
 - The Calculus of Probabilities
- 3 Conditional Probability and Independence
 - Bayes Theorem
 - Independence
- 4 Random Variables
- 5 Probability Function
 - Distribution Functions
- 6 Mass function
- 7 Reference

***Acknowledgement:** This slide is prepared based on Casella and Berger, 2002

Sample Space

ajfhgajd

Definition 1.1

The set of all possible outcomes of a particular experiment is called the *sample space* for the experiment, which generally denoted by S .

Example 1.1

In tossing a fair coin, we get either "heads" or "tails" , so the sample space is:

$$S = \{H, T\}.$$

Exercise:

- what is the sample space of a fair dice?
- Whats is the sample space of a credit risk problem ?
- Whats is the sample space of a Classification problem ?
- Define sample space of a user visiting a specific web page?
- Define sample space of a chat bot?

Definition 1.2

An event is any collection of possible outcomes of an experiment, which is, any subset of S (including S itself).

Exercise: pick any Amazon's product page, define several events.

- **Union:** The union of A and B is the set of elements that belong to **either A or B or both**.
- **Intersection:** The intersection of A and B is the set of elements that belong to **both** A and B.
- **Complement:** The complement of A is the set of all elements that are **not** in A.

$$A^c = \{x : x \notin A\}.$$

Definition 2.1

A collection of subsets S is called **sigma algebra (or Borel field)**, denoted by \mathbb{B} , if it satisfies the following three properties.

- 1 $\emptyset \in \mathbb{B}$ (the empty set is an element of \mathbb{B}).
- 2 If $A \in \mathbb{B}$, then $A^c \in \mathbb{B}$ (\mathbb{B} is closed under complement).
- 3 If $A_1, A_2, \dots \in \mathbb{B}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathbb{B}$.

Note that \emptyset is in the sigma algebra by (1). We know that $\emptyset^c = S$, so by (2), S is in \mathbb{B} .

Exercise:

- Is the set of all subsets of \mathbb{B} countable?
- Is \mathbb{B} a set?
- Define \mathbb{B} for a web page? Check if fits all 3 properties of sigma algebra.

if $A_1, A_2, \dots \in \mathbb{B}$ then $A_1^c, A_2^c, \dots \in \mathbb{B}$, by (2). Now, by (3), $\cup_{i=1}^{\infty} A_i^c \in \mathbb{B}$. However, using **De Morgan's Law**,

$$\left(\cup_{i=1}^{\infty} A_i^c \right)^c = \cap_{i=1}^{\infty} A_i.$$

- Is $\cap_{i=1}^{\infty} A_i$ countable or uncountable?

Then, by (2), $\cap_{i=1}^{\infty} A_i \in \mathbb{B}$ and \mathbb{B} is closed under countable intersections, as well. Note that sigma algebra is sometimes also denoted as σ – algebra and \mathbb{F} .

Hw: Prove that \mathbb{B} is closed under countable intersections.

Example 2.1

if S is finite or countable, then we can define for a given sample space S

$$\mathbb{B} = (\text{all subsets of } S \text{ including } S \text{ itself}).$$

- Take, for example, $S = \{1, 2, 3\}$. Then the sigma algebra defined in the previous bullet point is given by

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

- if S is uncountable, then it becomes difficult to describe \mathbb{B} . (why? **Exercise!**)

Definition 2.2

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

- 1 $P(A) \geq 0$ for all $A \in \mathbb{B}$.
- 2 $P(S) = 1$.
- 3 If $A_1, A_2, \dots \in \mathbb{B}$ are pairwise disjoint, then

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

Kolmogorov's Axioms

- These three points are usually referred to as the **Axioms of Probability** or the **Kolmogorov's Axioms**.
- Now, any function $P(\cdot)$ that satisfies the Kolmogorov Axioms is a valid probability function.

Example 2.2

In **tossing a fair coin**, we have, $\Omega = \{H, T\}$. The probability function is

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

as the coin is fair.

- observe that $\Omega = \{H\} \cup \{T\}$. Then, from Axiom 2 we must have

$$P(\{H\} \cup \{T\}) = 1.$$

- Since $\{H\}$ and $\{T\}$ are disjoint,

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

So,

$$P(\{H\} \cup \{T\}) = P(\{H\}) + P(\{T\}) = 1.$$

Axioms of Probability

- Our intuition and the Kolmogorov Axioms together tell us that $P(\{H\}) = P(\{T\}) = 1/2$.
- However, any non-negative probabilities that add up to one would have been valid, say $P(\{H\}) = 1/4$ and $P(\{T\}) = 3/4$. The reason we chose equal probabilities is our knowledge that the coin is fair!

Exercise: Use above example and provide a similar approach for a web page (buttons, option boxes, popup boxes,...)

The Foundation of Probabilities Functions

Theorem 2.1

If P is a probability function and A is any set in \mathbb{B} , then

- ① $P(\emptyset) = 0$ where \emptyset is the empty set.
 - ② $P(A) \leq 1$.
 - ③ $P(A^c) = 1 - P(A)$.
- **Proof: Exercise!**

The Foundation of Probabilities Functions

Theorem 2.2

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

- ① $P(B \cap A^c) = P(B) - P(A \cap B)$.
- ② $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- ③ If $A \subset B$, then $P(A) \leq P(B)$.

• **Proof: Exercise!**

Definition 3.1

If A and B are events in S , and $P(B) > 0$, then the **conditional probability** of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (1)$$

- In words, given that B has occurred, what is the probability that A will occur?
- By definition,

$$P(B|B) = 1,$$

as B has already occurred.

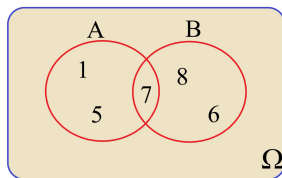
Conditional Probability

- If A and B are disjoint sets, $P(A \cap B) = P(\emptyset) = 0$ then:

$$P(A|B) = 0 = P(B|A)$$

- In fact, what happens in the conditional probability calculation is that B becomes the sample space.
- It is straightforward to verify that the probability function $P(.|B)$ satisfies Kolmogorov's Axioms, for any B for which $P(B) > 0$.

Conditional Probability



$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{7}{7+8+6} = \frac{1}{3}$$

Figure: Conditional Probability.

- Observe that since

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

we have

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A) \Rightarrow P(A|B) = P(B|A) \frac{P(A)}{P(B)} \quad (2)$$

- Which is known as Bayes's Rule.

Theorem 3.1

Let A_1, A_2, \dots be a partition of the sample, and let B be any set. Then, for each $i = 1, 2, \dots$

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)P(A_j)}.$$

This is actually not much more different than (2) since

$$\sum_{j=1}^{\infty} P(B|A_j)P(A_j) = \sum_{j=1}^{\infty} P(A_j \cap B) = P(B)$$

given that A_1, A_2, \dots is a partition of the sample space

Definition 3.2

Two events, A and B, are statistically independent if

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

Theorem 3.2

If A and B are independent events, then the following pairs are also independent:

- 1 A and B^c
- 2 A^c and B
- 3 A^c and B^c

Definition 4.1

A random variable is a function from a sample space S into the real numbers.

Experiment	Random Variable
Toss two dice	$X = \text{sum of the numbers}$
Toss a coin 10 times	$X = \text{number of heads in 10 tosses}$

- Suppose the sample space is $S = \{s_1, \dots, s_n\}$ and the original probability function is P . Define the new random variable

$$X : S \rightarrow X, \quad X = \{x_1, \dots, x_m\}$$

- Define the new probability function for X as P_X where

$$P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\}).$$

- P_X is an **induced probability function**, as it is defined in terms of the original probability function, P .
- If X is uncountable, the induced probability function is defined in a slightly different way. Namely, for any set $A \subset X$,

$$P_X(X \in A) = P(\{s \in S : X(s) \in A\}).$$

- In both cases, it is possible to show that the induced probability function satisfies the Kolmogorov Axioms.
- Note that the general convention in the literature is to assign capital letters to random variables and lower case letters to the particular value they take. Hence, for example, the number of people who answer "yes" in the survey could be X while a particular value, say 12, would be x .

Probability Function

- Suppose the sample space is $S = \{s_1, \dots, s_n\}$ and the original probability function is P . Define the new random variable

$$X : S \rightarrow X, \quad X = \{x_1, \dots, x_m\}$$

- Define the new probability function for X as P_X where

$$P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\}).$$

- P_X is an **induced probability function**, as it is defined in terms of the original probability function, P .
- If X is uncountable, the induced probability function is defined in a slightly different way. Namely, for any set $A \subset X$,

$$P_X(X \in A) = P(\{s \in S : X(s) \in A\}).$$

Probability Function

Example 5.1

Consider again the experiment of tossing a fair coin three times. Define the random variable X to be the number of heads obtained in the three tosses. A complete enumeration of the value of X for each point in the sample space is:

s	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(s)$	3	2	2	2	1	1	1	0

Probability Function

The range for the random variable X is $X = \{0, 1, 2, 3\}$. Assuming that all eight points s have probability $\frac{1}{8}$, by simply counting in the above display we see that the induced probability function on X is given by

x	0	1	2	3
$P_X(X=x)$	1/8	3/8	3/8	1/8

For example,

$$P_X(X = 1) = P(HTT, THT, TTH) = \frac{3}{8}$$

Distribution Functions

- All random variables are associated with a **distribution function**. This distribution function includes all information about the randomness of the variable.

Definition 5.1

The **cumulative distribution function** or **CDF** of a random variable X , denoted by $F_X(x)$, is defined by

$$F_X(x) = P_X(X \leq x), \quad \text{for all } x.$$

- When we write $P_X(X \leq x)$, we mean the probability that the random variable X takes a value equal to or smaller than x . The subscript X in $P_X(\cdot)$ denotes that this probability is obtained with respect to the probability distribution of X .

Distribution Functions

- In this particular case, this is too clear and so a bit redundant. However, if we consider $Y = f(X)$, then the notation provides clarification because we will have to deal with $P_X(Y \leq y)$.
- Note that the cdf is also generally denoted as simply the 'distribution.'

Example 5.2

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

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

- Note that, $F_X(x)$ is defined for all possible values of $x \in \mathcal{X}$. Hence,

$$P_X(x \leq 2.5) = P(X = 0, 1 \text{ or } 2) = 7/8.$$

Distribution Functions

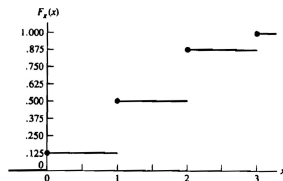


Figure: from Casella and Berger (2002, p.30). CDF of example 1.5.1

Distribution Functions

Theorem 5.1

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

- ① $\lim_{x \rightarrow -\infty} F_X(x) = 0$ and $\lim_{x \rightarrow +\infty} F_X(x) = 1$.
- ② $F_X(x)$ is a non-decreasing function of x .
- ③ $F_X(x)$ is right-continuous; that is, for every number x_0 , $\lim_{x \downarrow x_0} F_X(x) = F_X(x_0)$.

- We can also have a continuous cdf.

Definition 5.2

A random variable X is **continuous**(**discrete**) if $F_X(x)$ is a **continuous**(**step**) function of x .

Example 5.3

An example of a continuous cdf is the function

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

observe that

$$\lim_{x \rightarrow -\infty} F_X(x) = 0 \quad \text{since} \quad \lim_{x \rightarrow -\infty} e^{-x} = \infty,$$

$$\lim_{x \rightarrow \infty} F_X(x) = 1 \quad \text{since} \quad \lim_{x \rightarrow \infty} e^{-x} = 0,$$

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

where the final line proves that $F_X(x)$ is non-decreasing in x . Finally, by definition, $F_X(x)$ is right-continuous as it is continuous in the first place.

- This cdf is a special case of the [logistic distribution](#).

To be a cdf, a function has to possess some key properties.

Theorem 5.2

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

- 1 $\lim_{x \rightarrow -\infty} F_X(x) = 0$ and $\lim_{x \rightarrow +\infty} F_X(x) = 1$.
- 2 $F_X(x)$ is a non-decreasing function of x .
- 3 $F_X(x)$ is right-continuous; that is, for every number x_0 , $\lim_{x \downarrow x_0} F_X(x) = F_X(x_0)$.

Definition 5.3

The random variables X and Y are identically distributed if, for every set $A \in \mathbb{B}^1$, $P(X \in A) = P(Y \in A)$.

Theorem 5.3

The following two statements are equivalent:

- 1 The random variables X and Y are identically distributed.
- 2 $F_X(x) = F_Y(x)$ for every x .

Definition 6.1

The **probability mass function** of a discrete random variable is given by

$$f_X(x) = P(X = x) \text{ for all } x.$$

- **Notation convention:** for a given CDF F_X , the corresponding pdf is usually denoted by f_X , the corresponding lower-case letter.
- Similar to CDF being simply called the "distribution", the pdf is sometimes simply called the density.

Definition 6.2

The **probability density function** or pdf, $f_X(x)$, of a continuous random variable X is the function that satisfies

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

Using the Fundamental Theorem of Calculus, if f_X is continuous then

$$\frac{d}{dz} F_X(z)|_{z=x} = f_X(x).$$

Example 6.1

For the logistic distribution considered before, we have

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

and, hence,

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

Then, for continuous random variables in general,

$$\begin{aligned} P(a < X < b) &= F_X(b) - F_X(a) = \int_{-\infty}^b f_X(x) dx - \int_{-\infty}^a f_X(x) dx \\ &= \int_a^b f_X(x) dx. \end{aligned}$$

Theorem 6.1

A function $f_X(x)$ is a pdf (or pmf) of a random variable X if and only of

- 1 $f_X(x) \geq 0$ for all x .
- 2 $\sum_x f_X(x) = 1$ (discrete) or $\int_{-\infty}^{\infty} f_X(x) dx = 1$ (continuous).

- From a purely mathematical point of view, any non-negative function with a finite positive integral can be turned into a pdf or pmf. Take, for example, if

$$h(x) = \begin{cases} \geq 0 & \text{for } x \in A \\ 0 & \text{elsewhere} \end{cases}$$

and

$$\int_{x \in A} h(x) dx = K < \infty, \quad \text{where } K > 0,$$

then $f_X(x) = h(x)/K$ is a pdf of a random variable X taking values in A .

- In some cases, although $F_X(x)$ exists, $f_X(x)$ may not exist because $F_X(x)$ can be continuous but not differentiable. Therefore, sometimes statistical analysis would be based on $F_X(x)$ and not $f_X(x)$. We will not consider such cases here.

Casella, G., & Berger, R. (2002). *Statistical inference*. Cengage Learning.
<https://books.google.fr/books?id=FAUVEAAAQBAJ>