

Probability and Statistics

Omid Safarzadeh

February 9, 2022

Table of contents

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

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

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 Ω .

Example 1.1

In tossing two fair coins the sample space is:

$$\Omega = \{HH, HT, TH, TT\}.$$

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

The event space (\mathcal{A}) is the space of potential results of the experiment. In discrete case, \mathcal{A} is the power set of Ω

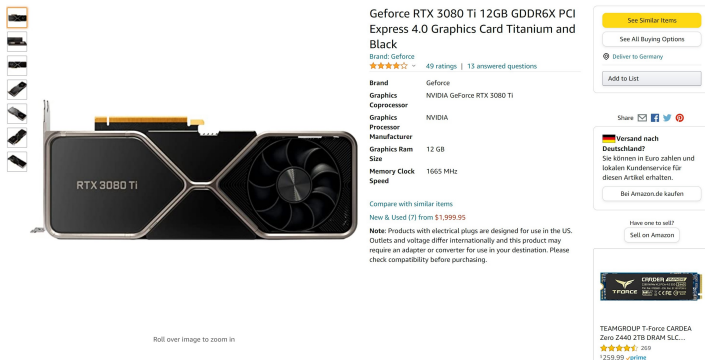
Definition 1.3

An event is any collection of possible outcomes of an experiment, which is, any subset of Ω .

(Ω, \mathcal{A}, P) is called Probability Space.

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

Example



GeForce RTX 3080 Ti 12GB GDDR6X PCI Express 4.0 Graphics Card Titanium and Black

Brand: GeForce
★★★★☆ 49 ratings | 13 answered questions

Brand	GeForce
Graphics Coprocessor	NVIDIA GeForce RTX 3080 Ti
Graphics Processor	NVIDIA
Manufacturer	
Graphics Ram Size	12 GB
Memory Clock Speed	1665 MHz

[Compare with similar items](#)
[New & Used \(7\) from \\$1,999.95](#)

Note: Products with electrical plugs are designed for use in the US. Outlets and voltage differ internationally and this product may require an adapter or converter for use in your destination. Please check compatibility before purchasing.

[See Similar Items](#)
[See All Buying Options](#)
[Deliver to Germany](#)
[Add to List](#)

Share [Email](#) [Facebook](#) [Twitter](#) [Pinterest](#)

Versand nach Deutschland?
Sie können in Euro zahlen und lokalen Kundenservice für diesen Artikel erhalten.
[Bei Amazon.de kaufen](#)

Have one to sell?
[Sell on Amazon](#)

TEAMGROUP T-Force CARDEA Zero Z440 2TB DRAM SLC...
★★★★☆ 209
\$259.99 [prime](#)

For example, in this web page, whole the web page is our sample space and all the clickable items, i.e. **Add to list**, **See Similar Items**, ..., are the events.

Definition 2.1

A collection, \mathbb{B} , of subsets Ω is called **sigma algebra**, if it satisfies the following:

- 1 $\emptyset \in \mathbb{B}$
- 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 $\bigcup_{i=1}^{\infty} A_i \in \mathbb{B}$.

The pair (\mathbb{B}, Ω) is called a measurable space or Borel Space.

Interpretation: Sigma Algebra is the collection of events that can be assigned probabilities.

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), $\cup_{i=1}^{\infty} A_i^c \in \mathbb{B}$. Use 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.

Example 2.1

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

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

- Take, for example, $\Omega = \{A, B, \dots, Z\}$. Then the sigma algebra is the power set of Ω .

Definition 2.2

Given a sample space Ω 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 $P(\Omega) = 1$.
- 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

Two events, A and B, are statistically independent if

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

Theorem 3.1

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

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.

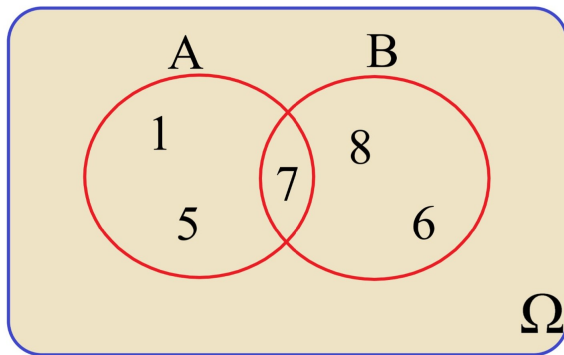
Conditional Probability

- 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



$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{7}{1+5+7+8+6}}{\frac{7+8+6}{1+5+7+8+6}} = \frac{\frac{7}{27}}{\frac{21}{27}} = \frac{1}{3}$$

Figure: Conditional Probability.

“You shall know a word by the company it keeps” (J. R. Firth 1957: 11)

Example 4.1

Omid is a _____ Data Scientist...

We want to predict _____ given other words!

$$P(w_{t+i}|w_t) = ? \quad (2)$$

For each position $t = 1, \dots, T$, predict context words within a window of fixed size m , given center word.

Ref: [Stanford CS224N course](#), Prof. Manning

Bayes Rule

- 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

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A) \Rightarrow P(A|B) = P(B|A) \frac{P(A)}{P(B)} \quad (3)$$

- Which is known as Bayes's Rule.

Theorem 4.1

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 (3) 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 5.1

A random variable is a function from a sample space Ω into the real numbers \mathbb{R} .

Experiment

Click on amazon product page
amazon product page visit

Random Variable

X = click through BUY Button .
 X = total number of calling the page
from server (GET API)

Random Variable example

Example 5.1

Consider tossing a fair coin three times. Define the random variable X to be the number of heads obtained in the three tosses. we have:

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

Random Variables in \mathbb{R}

Suppose the sample space is $\Omega = \{\omega_1, \dots, \omega_n\}$ and the original probability function is P . Define the new random variable

$$X : \Omega \rightarrow \mathbb{R}, \quad \text{take } A \in \mathbb{R}, A = \{x_1, \dots, x_m\}$$

$$P_X(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\}), A \in \mathbb{R}$$

Probability Function

- Suppose the sample space is $\Omega = \{\omega_1, \dots, \omega_n\}$ and the original probability function is P . Define the new random variable

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

- Define the new probability function for X as P_X where

$$P_X(X = x_i) = P(\{\omega_j \in \Omega : X(\omega_j) = x_i\}).$$

- 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 $A \subset \mathcal{X}$,

$$P_X(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\}).$$

Probability Function

- Exercise : show that induced probability function satisfies the Kolmogorov Axioms.
- we assign capital letters to random variables and lower case letters to the particular value they take.

Distribution Functions

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

Definition 6.1

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 .

Example 6.1

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 $x \in \mathbb{R}$. Hence,

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

Distribution Functions

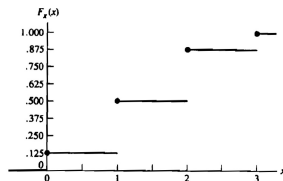


Figure: from Casella and Berger (2002, p.30). CDF of example 6.1

Distribution Functions

Theorem 6.1

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 6.2

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

Example 6.2

take Sigmoid function:

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

observe that

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

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

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

so $F_X(x)$ is non-decreasing in x .

$F_X(x)$ is continuous everywhere.

Distribution Functions

identically distributed

Theorem 6.2

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 6.3

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

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

- CDF F_X , then pdf is denoted by f_X
- the pmf is called "density". >>> for Discrete R.V.
- the pdf is also called "density" >>> for Continuous R.V.

Density and Mass Continuous Case

Definition 6.4

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

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

A note on notation : The expression " X has a distribution given by $F_X(x)$ " is abbreviated symbolically by " $X \sim F_X(x)$ ", where we read the symbol " \sim " as "is distributed as". We can similarly write $X \sim f_X(x)$ or, if X and Y have the same distribution, $X \sim Y$ ([Ref.](#))

Example 6.3

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 6.3

A function $f_X(x)$ is a pdf (/ 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).

Important for Simulating

- 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 $f_X(x)$.

Distribution of Functions of a Random Variable

- If X is a random variable with CDF F_X , then $Y = g(X)$ is also a random variable.
- Importantly, since Y is a function of X , we can determine its random behaviour in terms of the behaviour of X .
- Then, for any set A ,

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

This clearly shows that the distribution of Y depends on the function $g(\cdot)$ and the CDF F_X .

- Formally,

$$g(x) : \mathcal{X} \rightarrow \mathcal{Y},$$

where \mathcal{X} and \mathcal{Y} are the sample spaces of X and Y , respectively.

Distribution of Functions of a Random Variable

- Notice that the mapping $g(\cdot)$ is associated with the inverse mapping $g^{-1}(\cdot)$, a mapping from the subsets of \mathcal{Y} to those \mathcal{X} :

$$g^{-1}(A) = \{x \in \mathcal{X} : g(x) \in A\}. \quad (4)$$

- Therefore, the mapping $g^{-1}(\cdot)$ takes sets into sets, that is, $g^{-1}(A)$ is the **set of points in \mathcal{X} that $g(x)$ takes into the set A** .
- If $A = \{y\}$, a point set, then

$$g^{-1}(\{y\}) = \{x \in \mathcal{X} : g(x) = y\}.$$

- Now, if $Y = g(X)$, then for all $A \in \mathcal{Y}$,

$$\begin{aligned} P(Y \in A) &= P(g(X) \in A) \\ &= P(\{x \in \mathcal{X} : g(x) \in A\}) \\ &= P(X \in g^{-1}(A)), \end{aligned} \quad (5)$$

where the last line follows from (1). This defines the probability distribution of Y .

Distribution of Functions of a Random Variable

- The CDF of $Y = g(X)$ is

$$\begin{aligned}F_Y(y) &= P(Y \leq y) = P(g(X) \leq y) \\&= P(\{x \in \mathcal{X} : g(x) \leq y\}) \\&= \int_{x \in \mathcal{X} : g(x) \leq y} f_X(x) dx.\end{aligned}$$

Distribution of Functions of a Random Variable

Theorem 7.1

Let X have CDF $F_X(x)$, let $Y = g(X)$ and let \mathcal{X} and \mathcal{Y} be defined as

$$\mathcal{X} = \{x : f_X(x) > 0\} \quad \text{and} \quad \mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}. \quad (6)$$

- ① If g is an increasing function on \mathcal{X} , $F_Y(y) = F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$.
- ② If g is a decreasing function on \mathcal{X} and X is a continuous random variable, $F_Y(y) = 1 - F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$.

Proof: Exercise!

Distribution of Functions of a Random Variable

Theorem 7.2

Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is a monotone function. Let \mathcal{X} and \mathcal{Y} be defined as in (3). Suppose that $f_X(x)$ is continuous on \mathcal{X} and that $g^{-1}(y)$ has a continuous derivative on \mathcal{Y} . The pdf of Y is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & y \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases}$$

Proof: Exercise!

Distribution of Functions of a Random Variable

Example 7.1

Suppose $X \sim f_X(x) = 1$ for $0 < x < 1$ and 0 otherwise, which is the *uniform*(0,1) distribution. Observe that $F_X(x) = x$, $0 < x < 1$. We now make the transformation $Y = g(X) = -\log X$. Then,

$$g'(x) = \frac{d}{dx}(-\log x) = -\frac{1}{x} < 0 \quad \text{for } 0 < x < 1;$$

hence, $g(x)$ is monotone and has a continuous derivative on $0 < x < 1$. Also, $\mathcal{Y} = (0, \infty)$. Observe that $g^{-1}(y) = e^{-y}$. Then, using Theorem (7.2),

$$\begin{aligned} f_Y(y) &= 1 * |-e^{-y}| \quad \text{if } 0 < y < \infty \\ &= e^{-y} \quad \text{if } 0 < y < \infty. \end{aligned}$$

Theorem 7.3

Let X have continuous CDF $F_X(x)$ and define the random variable Y as $Y = F_X(X)$. Then, Y is uniformly distributed on $(0,1)$, that is

$$P(Y \leq y) = y, \quad 0 < y < 1.$$

• **Proof:** Exercise!

Distribution of Functions of a Random Variable

- This result connects any random variable with some CDF $F_X(x)$ with a uniformly distributed random variable. Hence, if we want to simulate random numbers from some distribution $F_X(x)$, all we have to do is to generate uniformly distributed random variables, Y , and then solve for $F_X(x) = y$. As long as we can compute $F_X^{-1}(y)$, we can generate random numbers from the distribution $F_X(x)$.



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