T2

Probability Theory and Machine Learning

Salvador Ruiz Correa

August 18, 2025

In probability theory, understanding how events relate to one another is essential for modeling uncertainty. The concept of independence describes situations where the occurrence of one event does not influence the likelihood of another. Building on this, conditional probability allows us to quantify the probability of an event given that another has occurred, forming the foundation for more complex reasoning. The chain rule extends this idea by expressing joint probabilities as a product of conditional probabilities, enabling the decomposition of multivariate distributions. Finally, Bayes' Theorem provides a powerful framework for updating beliefs in light of new evidence, reversing conditional probabilities to infer causes from observed outcomes. Together, these concepts form the backbone of probabilistic reasoning and inference.

Agenda:
1 Independence.
2 Conditional probability.
3 Chain rule.
4 Bayes Theorem.
5 Random Variables.

Independent Events

$A_m, A_n \in \mathscr{F}$ are independent $\implies P(A_m \cap A_n) = P(A_m)P(A_n)$

- $(A_n)_{n\in\mathbb{N}}$ are pair-wise independent $\implies A_m$ and A_n are independent for any $n \neq m$, $n, m \in \mathbb{N}$
- $(A_n)_{n\in I}$, are independent $\implies P(\bigcap_{n\in I} A_n) = \prod_{n\in I} P(A_n)$

Conditional probability

The conditional probability of event A conditioned to event B is defined as

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

where $P(B) > 0$. A conditional probability is a *probability measure*. If A and B are independent $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$.

Chain rule

Consider the events $A_1, A_2, \cdots, A_n$. The chain rule is defined as follows.

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) =$$
$$P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap \cdots \cap A_{n-1})$$

- Apply the chain rule to the following expression $P(A_1 \cap A_2 \cap A_3 \cap A_4)$.
- Is the following expression correct? $P(A_1 \cap A_2 \cap A_3) = P(A_1 \mid A_2 \cap A_3)P(A_2 \mid A_3)p(A_3)$.

Bayes Theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

where $P(B) > 0$. Recall that $P(B) = \sum_{n \in \mathbb{N}} P(B \cap A_n)$ with $\biguplus_{n \in \mathbb{N}} A_n = \Omega$. Therefore:

$$\begin{aligned} P(A_m \mid B) &= \frac{P(B \mid A_m)P(A_m)}{P(B)} \\ &= \frac{P(B \mid A_m)P(A_m)}{\sum_{n \in \mathbb{N}} P(B \cap A_n)} = \frac{P(B \mid A_m)P(A_m)}{\sum_{n \in \mathbb{N}} P(B \mid A_n)P(A_n)} \end{aligned}$$

Applying the Bayes Theorem

- Mr. Holmes now lives in Los Angeles. One morning when Homes leaves his house, he realizes that his grass is wet (H). Is it due to rain (R), or has he forgotten to turn off the sprinkler (S)? Next, he notices that the grass of his neighbor, Dr. Watson, is also wet (W).
  - What is the probability that Holme's grass is wet (H) given that he forgot to turn the sprinkler off (S)?
  - What is the probability that Holme's grass is wet (H) given that he forgot to turn the sprinkler off (S) and Dr. Watson's grass is also wet (W)?
- Consider the following events.
  - $A =$ "Holmes' grass is wet."
  - $A^c =$"Holmes grass is dry. "
  - $B =$"Holmes forgot to turn the sprinkler off."
  - $B' =$ "Holmes forgot to turn the sprinkler off and Watson's grass is wet."

  We write down the outcomes including these events and their corresponding probabilities in the tables shown below:

Ω	$P(\omega)$
$\omega_1 = \mathsf{H^cW^cS^cR^c}$	α_1
$\omega_2 = \mathsf{H^cW^cS^cR}$	α_2
$\omega_3 = \mathsf{H^cW^cSR^c}$	α_3
$\omega_4 = \mathsf{H^cW^cSR}$	α_4
$\omega_5 = \mathsf{H^cWS^cR^c}$	α_5
$\omega_6 = \mathsf{H^cWS^cR}$	α_6
$\omega_7 = \mathsf{H^cWSR^c}$	α_7
$\omega_8 = \mathsf{H^cWSR}$	α_8
$\omega_9 = \mathsf{HW^cS^cR^c}$	α_9
$\omega_{10} = \mathsf{HW^cS^cR}$	α_{10}
$\omega_{11} = \mathsf{HW^cSR^c}$	α_{11}
$\omega_{12} = \mathsf{HW^cSR}$	α_{12}
$\omega_{13} = \mathsf{HWS^cR^c}$	α_{13}
$\omega_{14} = \mathsf{HWS^cR}$	α_{14}
$\omega_{15} = \mathsf{HWSR^c}$	α_{15}
$\omega_{16} = \mathsf{HWSR}$	α_{16}

A	$P(\omega)$
$\omega_9 = \mathsf{HW^cS^cR^c}$	α_9
$\omega_{10} = \mathsf{HW^cS^cR}$	α_{10}
$\omega_{11} = \mathsf{HW^cSR^c}$	α_{11}
$\omega_{12} = \mathsf{HW^cSR}$	α_{12}
$\omega_{13} = \mathsf{HWS^cR^c}$	α_{13}
$\omega_{14} = \mathsf{HWS^cR}$	α_{14}
$\omega_{15} = \mathsf{HWSR^c}$	α_{15}
$\omega_{16} = \mathsf{HWSR}$	α_{16}

A^c	$P(\omega)$
$\omega_1 = \mathsf{H^cW^cS^cR^c}$	α_1
$\omega_2 = \mathsf{H^cW^cS^cR}$	α_2
$\omega_3 = \mathsf{H^cW^cSR^c}$	α_3
$\omega_4 = \mathsf{H^cW^cSR}$	α_4
$\omega_5 = \mathsf{H^cWS^cR^c}$	α_5
$\omega_6 = \mathsf{H^cWS^cR}$	α_6
$\omega_7 = \mathsf{H^cWSR^c}$	α_7
$\omega_8 = \mathsf{H^cWSR}$	α_8

B	$P(\omega)$
$\omega_3 = \mathsf{H^cW^cSR^c}$	α_3
$\omega_4 = \mathsf{H^cW^cSR}$	α_4
$\omega_7 = \mathsf{H^cWSR^c}$	α_7
$\omega_8 = \mathsf{H^cWSR}$	α_8
$\omega_{11} = \mathsf{HW^cSR^c}$	α_{11}
$\omega_{12} = \mathsf{HW^cSR}$	α_{12}
$\omega_{15} = \mathsf{HWSR^c}$	α_{15}
$\omega_{16} = \mathsf{HWSR}$	α_{16}

B'	$P(\omega)$
$\omega_7 = \mathsf{H^cWSR^c}$	α_7
$\omega_8 = \mathsf{H^cWSR}$	α_8
$\omega_{15} = \mathsf{HWSR^c}$	α_{15}
$\omega_{16} = \mathsf{HWSR}$	α_{16}

Box 1: Example solutions

We use conditional probability and the Bayes theorem to compute the requested probabilities.

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{\alpha_{11} + \alpha_{12} + \alpha_{15} + \alpha_{16}}{\sum_{k=9}^{16} \alpha_k}.$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{\alpha_{11} + \alpha_{12} + \alpha_{15} + \alpha_{16}}{\alpha_3 + \alpha_4 + \alpha_7 + \alpha_8 + \alpha_{11} + \alpha_{12} + \alpha_{15} + \alpha_{16}}.$$

$$P(B' \mid A) = \frac{P(B' \cap A)}{P(A)} = \frac{\alpha_{15} + \alpha_{16}}{\sum_{k=9}^{16} \alpha_k}.$$

$$P(A \mid B') = \frac{P(B' \mid A)P(A)}{P(B')} = \frac{\alpha_{15} + \alpha_{16}}{\alpha_7 + \alpha_8 + \alpha_{15} + \alpha_{16}}.$$

Random Variables

A random variable is also called a measurable function. Suppose you have two measurable spaces. One could be the pair of sample space and event space $(\Omega, \mathcal{F})$, and the other one could be some other arbitrary pair of a set and of its σ-algebra $(E, \mathcal{E})$, although usually, we choose $E = \mathbb{R}^n$ and $\mathcal{E} = \mathcal{B}(\mathbb{R}^n)$, $n = 1, 2, \ldots$. Then a measurable function is a function X that maps elements in Ω to elements in E with some additional properties. Notice how this function maps outcomes to elements of E, it does not map events (Figure 20).

The mapping X guarantees a correspondence between the events in our original event space and our transformed event space. For this reason, we require the random variable $X : \Omega \to E$ ($X : \Omega \to \mathbb{R}$) is such that to be such that the preimage $X^{-1}(B)$of any $\mathcal{E}$-measurable set $B \in \mathcal{B}$ is a $\mathcal{F}$ measurable set.

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F} \quad \forall B \in \mathcal{E}.$$

In the diagram below we can see how the set B, which is an element of $\mathcal{E}$ has a pre-image, $X^{-1}(B)$, which is an element of $\mathcal{F}$. From here on, we use $E = \mathbb{R}^n$, and $\mathcal{E} = \mathcal{B}(\mathbb{R}^n)$ to define a random variable.

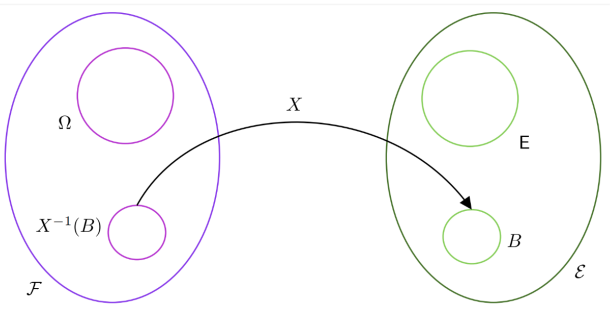


Figure 3: Random variable definition.

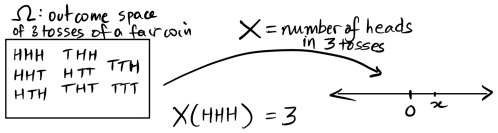


Figure 1: Random variable example defined on the real line $\mathbb{R}$. Notice that the inverse mapping X^{-1} maps real numbers into measurable events. For example $\{HHH\} = X^{-1}(3)$ and . $\{\{HHT\}, \{HTH\}, \{THH\}\} = X^{-1}(2)$ belong to the σ-algebra $\mathcal{F} = \mathcal{P}(\Omega)$.

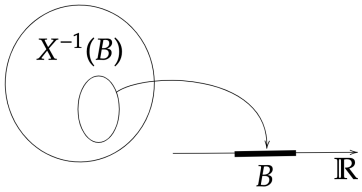


Figure 2: Random variable definition.

Definition 1: Random Variable

Let $(\Omega, \mathscr{F}, P)$ be a probability space. A *random variable* X is a $(\mathcal{F}/\mathcal{B}(\mathbb{R}^n))$ measurable map $X : \Omega \to \mathbb{R}^n$, if for every Borel set $B \in \mathcal{B}(\mathbb{R}^n)$:

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathscr{F}.$$

Clearly,

- $X^{-1}(B) \in \mathscr{F} \quad \forall B \in \mathcal{B}(\mathbb{R}^n)$.

Types of Random Variables

In practice, we seldom bother working with the abstract concept of a probability space $(\Omega, \mathscr{F}, P)$, but rather just focus on the distributional properties of a random variable X representing the random phenomenon. We are interested in random variables that are *discrete* and *continuous*.

- A random variable X is discrete if it only takes values on a countable set $\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$, which is called the support of X.
- A random variable X is a continuous random variable if it only takes values on a non-countable set Ω.
- Random variables can also be mixed.

Examples of Discrete Random Variables

- Consider the experiment in which we roll a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
  - Let $\mathscr{F} = \mathscr{P}(\Omega)$. The random variable $X_1(\omega) = \omega$, is $\mathscr{F}/\mathcal{B}(\mathbb{R})$ measurable. X_1 gives the exact outcome of the roll.
  - Let $\mathscr{F}_1 = \{\emptyset, \Omega, \{1, 3, 5\}, \{2, 4, 6\}\}$. The random variable

    $$X_2(\omega) = \begin{cases} 1, & \omega \in \{1, 3, 5\} \\ 0, & \omega \in \{2, 4, 6\} \end{cases}$$

    is $\mathscr{F}_1/\mathcal{B}(\mathbb{R})$ measurable; its value depends on whether the roll is odd or even.
  - If we only have information on whether the roll is odd or even, we can determine the value of X_2 but not the value of X_1.
  - The random variable X_1, is not $\mathscr{F}_1/\mathcal{B}(\mathbb{R})$ measurable. For instance, $X_1^{-1}(1) = \{1\} \notin \mathscr{F}_1$.
- Consider the experiment in which we roll a three-faced dice: $\Omega = \{-1, 0, 1\}$ and $\mathscr{F}_1 = \{\emptyset, \Omega, \{-1, 1\}, \{0\}\}$.
  - $X_1(\omega) = \omega$ is not $\mathscr{F}_1/\mathcal{B}(\mathbb{R})$ measurable. For example, $X^{-1}(1) = \{1\} \notin \mathscr{F}_1$.
  - $X_2(\omega) = \omega^2$ is $\mathscr{F}_1/\mathcal{B}(\mathbb{R})$ measurable.

- Let $(\Omega, \mathscr{F})$ describe throwing two fair dice, i.e. $\Omega := \{(i,k) : 1 \leq i,k \leq 6\}$, $\mathscr{F} = \mathscr{P}(\Omega)$. The total number of points thrown $X : \Omega \to \{2,3,\ldots,12\}$, $U((i,j)) = i + j$ is a measurable map.
- *A σ-algebra generated by a random variable U, denoted by $\sigma(U)$, is the smallest σ-algebra for which U is measurable. For example, for $\Omega = \{\mathsf{HH}, \mathsf{HT}, \mathsf{TH}, \mathsf{TT}\}$.

$$X_1(\omega) = \begin{cases} 1, & \omega \in \{\mathsf{HH}, \mathsf{HT}\}, \\ 0, & \omega \in \{\mathsf{TH}, \mathsf{TT}\}. \end{cases} \quad X_2(\omega) = \begin{cases} 2, & \omega \in \{\mathsf{HH}\}, \\ 1, & \omega \in \{\mathsf{HT}\}, \\ -1, & \omega \in \{\mathsf{TH}\}, \\ -2, & \omega \in \{\mathsf{TT}\}. \end{cases}$$

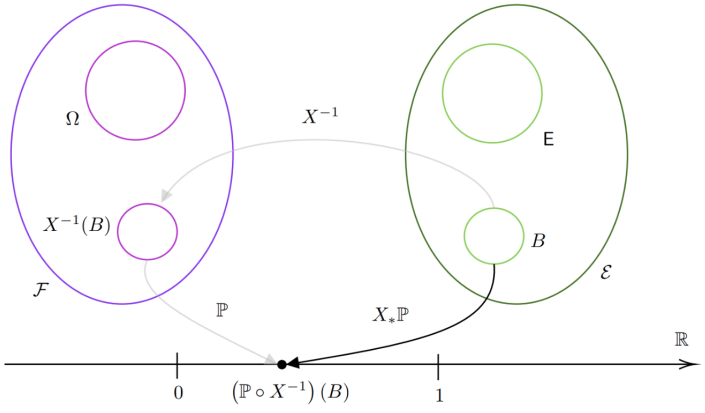
  $\sigma(X_1) = \{\emptyset, \{\mathsf{HH}, \mathsf{HT}\}, \{\mathsf{TH}, \mathsf{TT}\}, \Omega\}$ and $\sigma(X_2) = \mathscr{P}(\Omega)$. In particular, $\sigma(X_1) \subset \sigma(X_2) = \mathscr{P}(\Omega)$.

Probability Distributions

A probability distribution is also called a *push-forward* ($P_X = P_* X = P \circ X^{-1}$) measure of the *probability measure* P, via the random variable X. Suppose we have a probability space $(\Omega, \mathcal{F}, P)$ This means we can assign probabilities to events in $\mathcal{F}$. Now suppose we have a measurable space $(E, \mathcal{E})$ but we don't yet have a probability measure to measure events in it. How can we go about measuring sets in $\mathcal{E}$?

The key idea is that, given a set B in $\mathcal{E}$, we can use a random variable X to find the pre-image of such set in the event space $\mathcal{F}$ and then measure this set via the probability measure P. This will then be our probability measurement for B, as shown in the figure below.

The *probability distribution*, is defined as $P_X = P_* X = P \circ X^{-1} : \mathcal{E} \to [0,1]$. The probability distribution is therefore a function mapping sets in $\mathcal{E}$ into $[0,1]$. Here we use $(E, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, that is, $P \circ X^{-1} : \mathcal{B}(\mathbb{R}^n) \to [0,1]$.



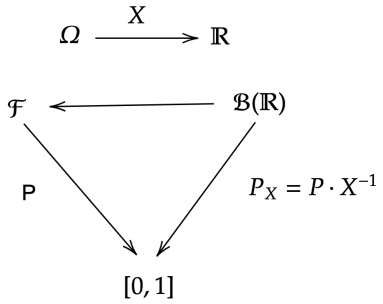


Figure 4: Probability distribution defined on the real line $\mathbb{R}$.

Figure 5: Probability distribution definition.

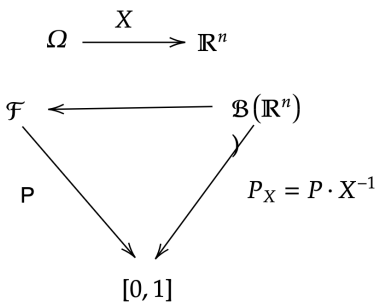


Figure 6: Probability distribution defined on the n-dimensional real space $\mathbb{R}$.

Definition 2: Probability Distribution

Let $(\Omega, P, \mathscr{F})$ be a measure space, and X a random variable $X : \Omega \to \mathbb{R}^n$, i.e. a measurable map. Then

$$P(X^{-1}(B)) = P(\{\omega : X(\omega) \in B\}) = P(X \in B).$$

is a probability measure called the *law* or *distribution* of the random variable X. Note: Here we use $(E, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, that is, $P \circ X^{-1} : \mathcal{B}(\mathbb{R}^n) \to [0, 1]$.

Examples of Probability Distributions for Discrete Random Variables

- Let $(\Omega, \mathscr{F})$ describe throwing two fair dice, i.e. $\Omega := \{(i,k) : 1 \leq i,k \leq 6\}$, $\mathscr{F} = \mathscr{P}(\Omega)$, and $P(\{i,j\}) = \frac{1}{36}$. The total number of points thrown $X : \Omega \to \{2,3,\ldots,12\}$, $X((i,j)) = i + j$ is a measurable map (Table 1 and Figure 21).

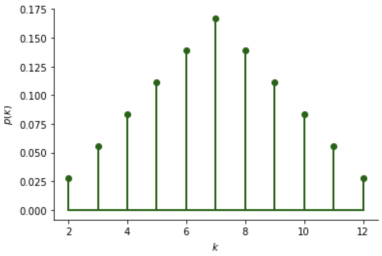


Figure 7: Probability distribution $P(K = k)$ in Table 1.

k	2	3	4	5	6	7	8	9	10	11	12
$P(X = k)$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

Table 1: The distribution of the random variable X.

- The *Bernoulli random* variable $X \in \{0,1\}$ with parameter p, $0 < \mathsf{p} \leq 1$ has the following probability distribution:

$$P(X = x \mid \mathsf{p}) := \mathsf{Bernoulli}(X = x \mid \mathsf{p}) = \mathsf{p}^x(1-\mathsf{p})^{1-x}.$$

  Hint: $P(X = 1 \mid \mathsf{p}) = \mathsf{p}$.

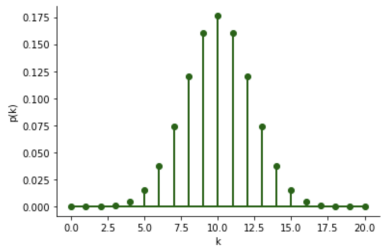


Figure 8: Binomial probability distribution function for $N = 20$ and $p = 0.5$.

- The *Binomial distribution* with parameters N and p is the discrete probability distribution of the number K of successes in a sequence of N independent Bernoulli trials (with parameter p). The probability distribution is

$$P(K = k \mid \mathsf{N}, \mathsf{p}) := \mathsf{Binomial}(K = k \mid \mathsf{p}, \mathsf{N}) = \binom{\mathsf{N}}{k} \mathsf{p}^k (1-\mathsf{p})^{N-k}$$

  for $k = 0, 1, 2, \ldots$ where

$$\binom{\mathsf{N}}{k} = \frac{\mathsf{N}!}{(\mathsf{N}-k)!k!}$$

  is the number of ways of choosing $K = k$ objects out of a total of N identical objects.

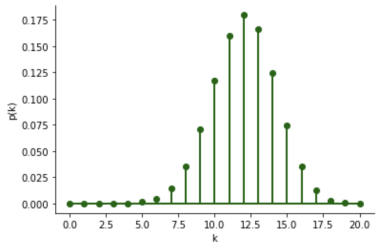


Figure 9: Binomial probability distribution function for $N = 20$ and $p = 0.6$.

Continuous Random Variables

A random variable X is a continuous random variable if there exists a non-negative function $f_X(\cdot)$ such that:

$$P(X \leq \omega) = \int_{-\infty}^{\omega} f_X(\alpha) d\alpha$$

for any $\omega \in \mathbb{R}$. The function f_X is called the probability density function of X. To simplify notation in practice we use $p(x) = f(x) = f_X(x)$. We remark on the abuse of notation.

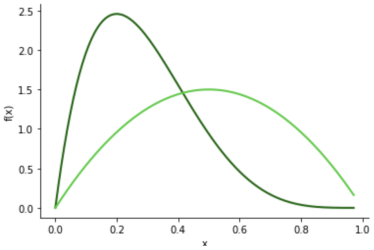


Figure 10: Beta probability density function. $\alpha = 2$, $\beta = 5$ (dark green), $\alpha = 2$, $\beta = 2$ (lime green).

Examples of Continuous Random Variables

- A random variable $M \in [0, 1]$ has a Beta distribution of variable with parameters α and β if the density function has the form

$$f_M(m \mid \alpha, \beta) := \text{Beta}(m \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} m^{\alpha-1}(1-m)^{\beta-1},$$

  where $\Gamma(x)$ is the Gamma function $\Gamma(x) = \int_0^x u^{x-1} e^{-u} du$.
- A random variable $X \in \mathbb{R}$ has a Gaussian or Normal distribution of variable with parameters μ and σ^2 if the density function has the form

$$f_X(x \mid \mu, \sigma^2) := \mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\},$$

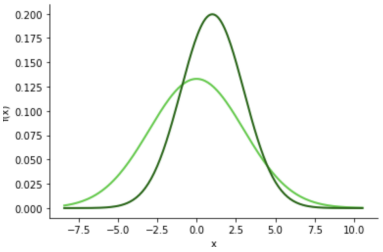


Figure 11: Gaussian probability density function. $\mu = 0$, $\sigma^2 = 3$ (dark green), $\mu = 2$, $\sigma^2 = 2$ (lime green).

Cummulative Distribution Function

For a random variable X, its cumulative distribution function (CDF) is defined as:

$$F_X(x) := P(X \leq x), \quad -\infty \leq x \leq \infty$$

Note that $P(X \leq x) = P \circ X^{-1}((-\infty, x])$, and:

- $F_X(x)$ is non-decreasing and right-continuous.
- $\lim_{x \to \infty} F_X(x) = 1$ and
- $\lim_{x \to -\infty} F_X(x) = 0$

Conversely, if a given function F_X satisfies the above properties, then it is a CDF of some random variable.

As an example, we show below the cumulative distribution of the random variable X in Table 1 (see Figures 21 and 22).

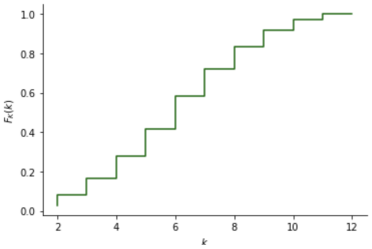


Figure 12: Cumulative distribution of the random variable X in Table 1.

k	2	3	4	5	6	7	8	9	10	11	12
$F_K(k)$	0.02	0.08	0.16	0.27	0.41	0.58	0.72	0.83	0.91	0.97	1