Probability Theory

SUZUKI, Atsushi

Probability theory

Probability theory handles random events, where the probability $\Pr(A) \in [0, 1]$ is defined for each event A.

Example

A	sunny	cloudy	rainy	others
$\Pr(A)$	0.4	0.2	0.3	0.1

An example simple weather forecast.

Here, the probability of the union of all the possible events is 1.

Random variables

When each elementary event is associated with a real value, then the set of those random events is called a ***random variable***.

One reason why we mainly consider random variables is that we can quantitatively discuss its random behavior.

Another important reason is that a computer only handles numeric values, so we need to associate each event with a value to handle them in a computer.

Discrete random variables

Definition

A random variable taking a value randomly in a discrete subset [1] of $\mathbb{R}$ (the set of real numbers) is called a ***discrete random variable***.
The subset of $\mathbb{R}$ in which a discrete random variable X takes a value is called the ***support*** or ***target space*** of X.

Example (Rolling an ideal six-sided dice)

Let X be the number that lands face-up when we roll an ideal six-sided dice.
The support of X is $\{1, 2, 3, 4, 5, 6\}$. The probability of each event is given by:

x	1	2	3	4	5	6
$\Pr(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Rolling an ideal six-sided dice

[1] Strictly speaking, "discrete" stands for "at most countable." Here, we say a set is at most countable if and only if there exists a surjective map from the set of integers to the set.

Probability mass function (PMF)

When we consider a univariate discrete random variable taking a value in a discrete set $\mathscr{X} = \{x_1, x_2, \dots\} \subset \mathbb{R}$, we can completely understand the behaviour of X by knowing the probability of X taking a value x, where $x \in \mathscr{X}$. Hence, we define a function describing those probabilities.

Definition (probability mass function (PMF))

Let X be a discrete random variable taking a value in a discrete set $\mathscr{X} \subset \mathbb{R}$. We define the ***probability mass function (PMF)*** $P_X : \mathscr{X} \to [0, 1]$ of the random variable X by

$$P_X(x) := \Pr(X = x). \tag{1}$$

Properties of a PMF

A PMF must satisfy the following:

- **(Nonnegativity)** $P_X(x) \geq 0$ for all $x \in \mathcal{X}$.
- **(The sum)** $\sum_{x \in \mathcal{X}} P_X(x) = 1$.

PMF tells us all we want to know.

If we want to know, for example, $\Pr(a \le X \le b)$, we can find it by the PMF:

$$\Pr(a \le X \le b) = \sum_{a \le x \le b} P_X(x). \tag{2}$$

Example (Rolling an ideal six-sided dice)

Let X be the number that lands face-up when we roll an ideal six-sided dice.
The support of X is $\{1,2,3,4,5,6\}$. The PMF of each event is given by:

x	1	2	3	4	5	6
$P_X(x) := \Pr(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Rolling an ideal six-sided dice

Here, $\Pr(2 \le x \le 4)$ is given by
$\sum_{2 \le x \le 4} P_X(x) = P_X(2) + P_X(3) + P_X(4) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$.

A frequency is a discrete random variable

Suppose that we have m data points taking values in $\mathbb{R}$. If we sample a data point uniform-randomly, the value of the data point is a discrete random variable.

The probability distribution of the random variable constructed from the data points this way is called the ***frequency*** or ***empirical distribution***.

Example (Exam results)

Suppose that we have $m = 20$ students and consider their results in an exam. For $x \in \mathcal{X} = \{0, 1, 2, 3, 4, 5\}$, we denote the number of the students who got a score x by m_x. Let X be the score of the student sampled uniform-randomly from the 20 students. The probability $\Pr(X = x)$ equals to $\frac{m_x}{m}$. For example,

Score x	0	1	2	3	4	5
# students m_x	3	2	3	5	6	1
$P_X := \Pr(X = x) = \frac{m_x}{m}$	0.15	0.10	0.15	0.25	0.30	0.5

Exam result data points and the frequency.

A function of a random variable

Let $f : \mathbb{R} \to \mathbb{R}$ be a function and X be a random variable.

If we input X to f, the return value $f(X)$ is also a random variable.

Let's find its PMF $P_{f(X)}$.

A function of a random variable: Example

Define f by $f(x) = x^2$, and the PMF P_X is given by the following table.

x	-1	0	+1
$P_X(x)$	0.2	0.3	0.5

Example random function and its PMF.

We want to find the PMF of $f(X) = X^2$ denoted by $P_{f(X)}$ or P_{X^2}.

By definition $P_{X^2}(0) = \Pr\left(X^2 = 0\right)$.

Since $X^2 = 0 \Leftrightarrow X = 0$ holds,[2] we have that $\Pr\left(X^2 = 0\right) = \Pr(X = 0) = 0.3$.

This case is easy since only one value of X corresponds to $X^2 = 0$.

[2] The symbol $\Leftrightarrow$ indicates a necessary and sufficient condition, or equivalence.

A function of a random variable: Example

Define f by $f(x) = x^2$, and the PMF P_X is given by the following table.

x	-1	0	+1
$P_X(x)$	0.2	0.3	0.5

Example random function and its PMF.

We want to find the PMF of $f(X) = X^2$ denoted by $P_{f(X)}$ or P_{X^2}.

By definition $P_{X^2}(1) = \Pr\left(X^2 = 1\right)$.

Since "$X^2 = 1$" $\Leftrightarrow$ "$X = -1$ or $X = +1$" holds, we have that

$$\begin{aligned} \Pr\left(X^2 = 1\right) &= \Pr(\text{“}X = -1 \text{ or } X = -1\text{”}) \\ &= \Pr(X = -1) + \Pr(X = +1) = 0.2 + 0.5 = 0.7, \end{aligned} \tag{3}$$

where the second equation comes from the sum law[2].

[2] The sum law applies because "$X = -1$ and $X = -1$" do not happen at the same time.

A function of a random variable: Example

Define f by $f(x) = x^2$, and the PMF P_X is given by the following table.

x	-1	0	+1
$P_X(x)$	0.2	0.3	0.5

Example random function and its PMF.

We want to find the PMF of $f(X) = X^2$ denoted by $P_{f(X)}$ or P_{X^2}.

To wrap up,

x	0	+1
$P_{X^2}(x)$	0.3	0.7

The PMF of X^2.

A function of a random variable

In general, let X is a random variable and $f : \mathbb{R} \to \mathbb{R}$ is a function.

We can find the PMF of $f(X)$ as follows.

$$P_{f(X)}(y) = \sum_{x:f(x)=y} P_X(x), \tag{3}$$

where $\sum_{x:(condition)}$ means we sum up the term if x satisfies the condition.

Note: Although functions of a random variable are important in practice, the above formula is less important since we are not often interested in the PMF of the return value. We are more interested in the return value's ***expectation***, which can be calculated without the above formula.

Summary statistics

Motivation: A probability mass function might have too much information to understand the behaviour of a random variable intuitively.

Hence, we often want to calculate a single value (or a few values) that describes a distribution, called a ***descriptive statistic*** or ***summary statistic***[2].

[2] These words are often used to distinguish them from inferential statistics.

Summary statistics: examples

Central tendency measures give a representative value of the values that the random variable takes, e.g., ***expectation***, ***median***, ***mode***, etc.

Variability measures show how spread values the random variable takes, e.g., ***range***, ***variance***, ***standard deviation***, ***quartile deviation***.

Other measures e.g., kurtosis, skewness.

Central tendency measure 1: Expectation (mean)

Definition (Expectation)

The ***expectation*** of a discrete random variable X, denoted by $\mathbb{E}X$, $\mathbf{E}X$, $\langle X \rangle$, or $\overline{X}$, is the weighted mean of the values with the probability masses as weights. That is

$$\mathbb{E}X := \sum_{x \in \mathscr{X}} x P_X(x). \tag{4}$$

The expectation is also called the ***mean***.

Central tendency measure 2: Median

If a distribution takes some extremely large or small values, the expectation is significantly influenced by the probability of the random variable taking such values.

In such cases, some might want to use the ***median*** as a summary statistic. Roughly speaking, the median is defined so that the random variable is larger than the median in 50% probability and smaller than the median in 50% probability. The strict definition of the median is somewhat technical, so it is more important to understand intuition and how to calculate it.

Central tendency measure 2: Median (cont.)

Definition (The broader definition of the median)

Let $P : \mathbb{R} \to [0, 1]$ be the probability mass function of a univariate discrete random variable X. If a real value $m \in \mathbb{R}$ satisfies the following equation, then m is called a ***median*** of the distribution of X:

$$\Pr(X \leq m) \geq \frac{1}{2} \text{ and } \Pr(X \geq m) \geq \frac{1}{2}. \tag{5}$$

We can often see the above definition in the context of probability theory.

Central tendency measure 2: Median (cont.)

Definition (The narrower definition of the median)

Let $P : \mathbb{R} \to [0, 1]$ be the probability mass function of a univariate discrete random variable X. If a real value $m \in \mathbb{R}$ satisfies the following equation, then m is called a ***median*** of the distribution of X:

$$\Pr(X \leq m) \geq \frac{1}{2} \text{ and } \Pr(X \geq m) \geq \frac{1}{2}. \tag{6}$$

We can often see the above definition in the context of probability theory.

Expectation of a function

If X is a random variable and f is a function, $f(X)$ is again a random variable. Hence, we can define the expectation of $f(X)$.

The expectation $\mathbb{E}f(X)$ often gives us important information as well as the original expectation $\mathbb{E}X$. The most important example is the ***variance*** of a random variable, which is the most frequently used variability measure.

Variability measure: Variance

Variability measures show how much the random variable deviates from the “center”.

The most representative one is the ***variance***, defined based on the ***square deviation***.

Let X be a random variable and μ be its expectation. The ***square deviation*** of X is defined as $\left(X-\mu\right)^2$. If X is far (whether large or not) from μ, the square deviation $\left(X-\mu\right)^2$ is large. Hence, we can regard its expectation as a variability measure. This is the idea of the variance.

Definition (Variance)

Let X be a random variable and assume that the expectation $\mu := \mathbb{E}X$ exists. Then, the ***variance*** $\mathbb{V}[X] \in \mathbb{R}_{\geq 0}$ is defined as the expectation of the squared deviation [3] $\left(X-\mu\right)^2$, that is,

$$\mathbb{V}[X] := \mathbb{E}\left(X-\mu\right)^2. \tag{7}$$

[3] One reason for considering the square is to ignore the sign. For the same reason, the expectation of the absolute deviation is also used. However, the variance, the expectation of the squared deviation, is much more often used owing to the central limit theorem.

Calculating the expectation of a function

The variance of a random variable X is $\mathbb{E}\, f(X)$ where f is defined by $f(x) = \big(x - \mu\big)$. In particular, the variance is the expectation of the return value $f(X)$. How can we calculate it?

In theory, we can calculate it by finding the PMF $f(X)$.

Nevertheless, we can skip this step by the following theorem.

Theorem

Let X be a discrete random variable, whose support is $\mathcal{X}$ and P_X be its PMF. Also, let f be a function. The expectation $\mathbb{E}\, f(X)$ is given by

$$\mathbb{E}\, f(X) = \sum_{x \in \mathcal{X}} f(x) P_X(x). \tag{8}$$

Calculating the variance

In particular, we can calculate the variance of a discrete random variable as follows.

Theorem

Suppose that X is a discrete random variable taking values in $\mathscr{X} \subset \mathbb{R}$ and $\mu := \mathbb{E}X$ is its expectation.
The variance $\mathbb{V}[X]$ is given by

$$\mathbb{V}[X] = \sum_{x \in \mathscr{X}} \big(x - \mu\big)^2 P_X(x). \tag{9}$$

Variability measure: Standard deviation

Variance's interpretation is somewhat tricky since it is not "linear." Specifically, the variance of $10X$ is 100 times as large as that of X.

To make it "linear", we consider the square root of the variance, called the ***standard deviation*** of the random variable.

Definition (Standard deviation)

The ***standard deviation*** $\sigma[X] \in \mathbb{R}$ of the random variable X is defined as

$$\sigma[X] := \sqrt{\mathbb{V}[X]}. \tag{10}$$

As expected, $\sigma[cX] = |c|\sigma[X]$ for $c \in \mathbb{R}$.

Multiple random variables

Example

- The prices of multiple stocks.
- The pixels of an image.
- The values at each time frame in a wave file.

When we consider multiple random variables, knowing each probability mass function is not sufficient to know their stochastic behavior completely.

Knowing multiple random variables $\neq$ knowing multiple PMFs

If we have two discrete random variables X and Y, then just knowing each probability mass function is not sufficient.

Rather, what we need to know is the distribution of the **pair** (X, Y), which is called the ***joint distribution*** of the random variables X and Y.

Joint distribution and marginal distribution

In general, the ***joint distribution*** refers to the distribution of the tuple of multiple random variables. For example, if we have two random variables X and Y, the joint distribution refers to the distribution of the pair (X, Y).

In contrast, when we consider multiple random variables, the distribution of a single random variable is called the ***marginal distribution*** of the random variable to distinguish it from the joint distribution.

Joint probability mass function (two variable cases)

If we have two discrete random variables X and Y, then just knowing each probability mass function is not sufficient. Rather, what we need to know is the probability of the pair (X, Y) taking every pair of values $(x, y) \in \mathcal{X} \times \mathcal{Y}$. That is, the following ***joint probability mass function (joint PMF)*** has all the information that we need.

Definition (two-variable Joint PMF)

Let X and Y be discrete random variables taking a value in discrete sets $\mathcal{X}$ and $\mathcal{Y}$, respectively, where $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$. We define the ***joint probability mass function (joint PMF)*** $P_{X,Y} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ of the pair of random variables X, Y by

$$P_{X,Y}(x, y) := \Pr(X = x, Y = y). \tag{11}$$

Joint PMF example

The random variables X and Y are the scores of a math test and a history test, respectively, where we uniform-randomly sample a student.

		x				$P_Y(y)$
		0	1	2	3	
y	0	0.16	0.04	0.02	0.06	0.28
	1	0.18	0.04	0.04	0.16	0.42
	2	0.06	0.02	0.08	0.14	0.30
$P_X(x)$		0.40	0.10	0.14	0.36	

An example of $P_{X,Y}(x, y) := \Pr(X = x \wedge Y = y)$

Joint probability mass function (general cases)

If we have m discrete random variables $X_1, X_2, \ldots, X_m$, then all we need to know is the following joint PMF.

Definition (Joint PMF (general cases))

Let $X_1, X_2, \ldots, X_m$ be discrete random variables taking a value in discrete sets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_m \subset \mathbb{R}$, respectively. We define the ***joint probability mass function (joint PMF)*** $P_{X_1, X_2, \ldots, X_m} : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_m \to [0, 1]$ of random variables $X_1, X_2, \ldots, X_m$ by

$$P_{X_1, X_2, \ldots, X_m}(x_1, x_2, \ldots, x_m) \coloneqq \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m). \tag{12}$$

Marginal PMF (two variable cases)

The joint PMF can tell us the PMFs of each discrete random variable, called ***marginal PMF***. For two discrete random variables X and Y that takes a value in $\mathscr{X}$ and $\mathscr{Y}$, respectively, suppose that the joint PMF is $P_{X,Y} : \mathscr{X} \times \mathscr{Y} \to [0,1]$. Then, the marginal PMFs P_X and P_Y are given by

$$P_X(x) = \sum_{y \in \mathscr{Y}} P_{X,Y}(x,y), \quad P_Y(y) = \sum_{x \in \mathscr{X}} P_{X,Y}(x,y), \tag{13}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
y	0	0.16	0.04	0.02	0.06	0.28
	1	0.18	0.04	0.04	0.16	0.42
	2	0.06	0.02	0.08	0.14	0.30
$P_X(x)$		**0.40**	0.10	0.14	0.36	

An example of $P_{X,Y}(x,y) \coloneqq \Pr(X = x \wedge Y = y)$

In the above example, we can calculate the marginal PMF from the joint PMF as follows.

$$\begin{aligned} P_X(0) &= P_{X,Y}(0,0) + P_{X,Y}(0,1) + P_{X,Y}(0,2) \\ &= 0.16 + 0.18 + 0.06 = \mathbf{0.40}. \end{aligned} \tag{14}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
y	0	0.16	0.04	0.02	0.06	0.28
	1	0.18	0.04	0.04	0.16	0.42
	2	0.06	0.02	0.08	0.14	0.30
$P_X(x)$		0.40	**0.10**	0.14	0.36	

An example of $P_{X,Y}(x,y) := \Pr(X = x \wedge Y = y)$

In the above example, we can calculate the marginal PMF from the joint PMF as follows.

$$\begin{aligned} P_X(1) &= P_{X,Y}(1,0) + P_{X,Y}(1,1) + P_{X,Y}(1,2) \\ &= 0.04 + 0.04 + 0.02 = \mathbf{0.10}. \end{aligned} \tag{14}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
y	0	0.16	0.04	0.02	0.06	0.28
	1	0.18	0.04	0.04	0.16	0.42
	2	0.06	0.02	0.08	0.14	0.30
$P_X(x)$		0.40	0.10	**0.14**	0.36	

An example of $P_{X,Y}(x,y) \coloneqq \Pr(X = x \wedge Y = y)$

In the above example, we can calculate the marginal PMF from the joint PMF as follows.

$$\begin{aligned} P_X(2) &= P_{X,Y}(2,0) + P_{X,Y}(2,1) + P_{X,Y}(2,2) \\ &= 0.02 + 0.04 + 0.08 = \mathbf{0.14}. \end{aligned} \tag{14}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
y	0	0.16	0.04	0.02	0.06	0.28
	1	0.18	0.04	0.04	0.16	0.42
	2	0.06	0.02	0.08	0.14	0.30
$P_X(x)$		0.40	0.10	0.14	**0.36**	

An example of $P_{X,Y}(x,y) \coloneqq \Pr(X = x \wedge Y = y)$

In the above example, we can calculate the marginal PMF from the joint PMF as follows.

$$\begin{aligned} P_X(3) &= P_{X,Y}(3,0) + P_{X,Y}(3,1) + P_{X,Y}(3,2) \\ &= 0.06 + 0.16 + 0.14 = \mathbf{0.36}. \end{aligned} \tag{14}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.16	0.04	0.02	0.06	**0.28**
y	1	0.18	0.04	0.04	0.16	0.42
	2	0.06	0.02	0.08	0.14	0.30
$P_X(x)$		0.40	0.10	0.14	0.36	

An example of $P_{X,Y}(x,y) := \Pr(X = x \wedge Y = y)$

In the above example, we can calculate the marginal PMF from the joint PMF as follows.

$$\begin{aligned} P_Y(0) &= P_{X,Y}(0,0) + P_{X,Y}(1,0) + P_{X,Y}(2,0) + P_{X,Y}(3,0) \\ &= 0.16 + 0.04 + 0.02 + 0.06 = \mathbf{0.28}. \end{aligned} \tag{14}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
y	0	0.16	0.04	0.02	0.06	0.28
	1	0.18	0.04	0.04	0.16	**0.42**
	2	0.06	0.02	0.08	0.14	0.30
$P_X(x)$		0.40	0.10	0.14	0.36	

An example of $P_{X,Y}(x,y) := \Pr(X = x \wedge Y = y)$

In the above example, we can calculate the marginal PMF from the joint PMF as follows.

$$\begin{aligned} P_Y(1) &= P_{X,Y}(0,1) + P_{X,Y}(1,1) + P_{X,Y}(2,1) + P_{X,Y}(3,1) \\ &= 0.18 + 0.04 + 0.04 + 0.16 = \mathbf{0.42}. \end{aligned} \tag{14}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
y	0	0.16	0.04	0.02	0.06	0.28
	1	0.18	0.04	0.04	0.16	0.42
	2	0.06	0.02	0.08	0.14	**0.30**
$P_X(x)$		0.40	0.10	0.14	0.36	

An example of $P_{X,Y}(x,y) := \Pr(X = x \wedge Y = y)$

In the above example, we can calculate the marginal PMF from the joint PMF as follows.

$$\begin{aligned} P_Y(2) &= P_{X,Y}(0,2) + P_{X,Y}(1,2) + P_{X,Y}(2,2) + P_{X,Y}(3,2) \\ &= 0.06 + 0.02 + 0.08 + 0.14 = \mathbf{0.30}. \end{aligned} \tag{14}$$

Conditional distribution

If two random variables are “related,” then we get more precise information about a random variable’s distribution by knowing the value of the other random variable.

The ***conditional distribution*** is a piece of such information.

The conditional information is the distribution of one variable

Conditional distribution example

Suppose that the joint PMF of random variables X and Y is given by:

		x				$P_Y(y)$
		0	1	2	3	
y	0	0.16	0.04	0.02	0.06	0.28
	1	0.18	0.04	0.04	0.16	0.42
	2	0.06	0.02	0.08	0.14	0.30
$P_X(x)$		0.40	0.10	0.14	0.36	

An example of $P_{X,Y}(x, y) := \Pr(X = x \wedge Y = y)$

Conditional distribution example

Suppose that the joint PMF of random variables X and Y is given by:

		x				$P_Y(y)$
		0	1	2	3	
y	0	0.16	0.04	0.02	0.06	0.28
	1	0.18	0.04	0.04	0.16	0.42
	2	0.06	0.02	0.08	0.14	0.30
$P_X(x)$		0.40	0.10	0.14	0.36	

An example of $P_{X,Y}(x, y) := \Pr(X = x \wedge Y = y)$

If we know $Y = 0$, then the probability masses of X is proportional to the joint masses.

Conditional distribution example

If we know $Y = 0$, then the probability masses of X is proportional to the joint masses.

		x				$P_Y(y)$
		0	1	2	3	
y	0	0.16	0.04	0.02	0.06	0.28

The joint probabilities $P_{X,Y}(x, y) := \Pr(X = x \wedge Y = y)$ where $Y = 0$.

We want to find the ***conditional probability*** $P_{X|Y}(x|0)$, which indicates the probability of "$X = x$" when we know $Y = 0$.

The sum $P_{X|Y}(0|0) + P_{X|Y}(1|0) + P_{X|Y}(2|0) + P_{X|Y}(3|0)$ of the conditional probabilities must be 1 for them to be probabilities.

Hence, the conditional probability $P_{X|Y}(x|0)$ is each joint probability over the sum, i.e.,

$$P_{X|Y}(x|0) = \frac{P_{X,Y}(x,0)}{P_{X,Y}(0,0) + P_{X,Y}(1,0) + P_{X,Y}(2,0) + P_{X,Y}(3,0)} = \frac{P_{X,Y}(x,0)}{P_Y(0)}. \quad (15)$$

Conditional distribution calculation example

the conditional probability $P_{X|Y}(x|0)$ is each joint probability over the sum, i.e.,

$$P_{X|Y}(x|0) = \frac{P_{X,Y}(x,0)}{P_{X,Y}(0,0)+P_{X,Y}(1,0)+P_{X,Y}(2,0)+P_{X,Y}(3,0)} = \frac{P_{X,Y}(x,0)}{P_Y(0)}. \quad (16)$$

Specifically, we can calculate the conditional probabilities as follows:

			x				$P_Y(y)$
			0	1	2	3	
y	0	The joint probabilities $P_{X,Y}(x,y)$	0.16	0.04	0.02	0.06	0.28

The joint probabilities $P_{X,Y}(x,y)$ and conditional probabilities $P_{X|Y}(x|y)$ where $Y=0$.

When $Y=0$, the probability of $X=0$ or 1 is larger than when we have no information

Conditional distribution calculation example

the conditional probability $P_{X|Y}(x|0)$ is each joint probability over the sum, i.e.,

$$P_{X|Y}(x|0) = \frac{P_{X,Y}(x,0)}{P_{X,Y}(0,0)+P_{X,Y}(1,0)+P_{X,Y}(2,0)+P_{X,Y}(3,0)} = \frac{P_{X,Y}(x,0)}{P_Y(0)}. \quad (16)$$

Specifically, we can calculate the conditional probabilities as follows:

			x				$P_Y(y)$
			0	1	2	3	
y	0	The joint probabilities $P_{X,Y}(x,y)$	0.16	0.04	0.02	0.06	0.28
		The conditional probabilities $P_{X\|Y}(x\|y)$	$\frac{0.16}{0.28}$	$\frac{0.04}{0.28}$	$\frac{0.02}{0.28}$	$\frac{0.06}{0.28}$	

The joint probabilities $P_{X,Y}(x,y)$ and conditional probabilities $P_{X|Y}(x|y)$ where $Y=0$.

When $Y=0$, the probability of $X=0$ or 1 is larger than when we have no information

Conditional distribution calculation example

the conditional probability $P_{X|Y}(x|0)$ is each joint probability over the sum, i.e.,

$$P_{X|Y}(x|0) = \frac{P_{X,Y}(x,0)}{P_{X,Y}(0,0)+P_{X,Y}(1,0)+P_{X,Y}(2,0)+P_{X,Y}(3,0)} = \frac{P_{X,Y}(x,0)}{P_Y(0)}. \quad (16)$$

Specifically, we can calculate the conditional probabilities as follows:

			x				$P_Y(y)$
			0	1	2	3	
y	0	The joint probabilities $P_{X,Y}(x,y)$	0.16	0.04	0.02	0.06	0.28
		The conditional probabilities $P_{X\|Y}(x\|y)$	0.57	0.14	0.07	0.21	
		The marginal distribution $P_X(x)$	0.40	0.10	0.14	0.36	

The joint probabilities $P_{X,Y}(x,y)$ and conditional probabilities $P_{X|Y}(x|y)$ where $Y=0$.

When $Y=0$, the probability of $X=0$ or 1 is larger than when we have no information

Independency of random variables

Suppose that the conditional PMF always equals to the marginal PMF, i.e., $P_{X|Y}(x|y) = P_X(x)$ for all x and y.

It means that Y has no relation to X. In this case, we say that X and Y are ***independent***.

Definition

Let X and Y be discrete random variables. If one of the following equivalent conditions[4] holds, we say that X and Y.

- $P_{X|Y}(x|y) = P_X(x)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.
- $P_{Y|X}(y|x) = P_Y(y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.
- $P_{X,Y}(x,y) = P_X(x)P_Y(y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

[4] Specifically, if one condition holds, then the other two conditions also hold.

Conditional probability calculation

In general, we can calculate the conditional PMF from the joint PMF and the marginal PMF as follows:

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}. \tag{17}$$

Since we can calculate the marginal probability $P_Y(y)$ by $P_Y(y) = \sum_{x \in \mathcal{X}} P_{X,Y}(x,y)$ using the joint PMF $P_{X,Y}$, we can calculate the conditional PMF only from the joint PMF in theory.

Summary statistics for multiple RVs: Covariance

One principal question about the relation between two random variables X and Y is: "Do they tend to take large values simultaneously, or does one tend to be small when the other is large?"

To answer the question, we consider the product of $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$. $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$ are positive if X and Y are relatively large, respectively.

Hence, if X and Y tend to take large values simultaneously, then the product $(X - \mathbb{E}X)(Y - \mathbb{E}Y)$ tend to be positive.

Conversely, if one tends to be small when the other is large, then the product $(X - \mathbb{E}X)(Y - \mathbb{E}Y)$ tend to be negative.

The above observation leads us to the definition of the ***covariance***.

Definition of the covariance

Definition

Let X and Y be random variables. Then, the ***covariance*** $\mathrm{Cov}(X,Y) \in \mathbb{R}$ between the two random variables X and Y is defined by

$$\mathrm{Cov}(X,Y) := \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]. \tag{18}$$

A positive covariance indicates that the two random variables tend to take relatively large values simultaneously. A negative covariance indicates that when one of the two takes a relatively large value, then the other tend to take a relatively small value.

Correlation

The covariance considers the scale of each random variable, not only the relation between them. Specifically, for $a, b \in \mathbb{R}$, we have that

$$\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X, Y). \tag{19}$$

This implies that just multiplying the random variables by some factors changes the value of the correlation although the relation between aX and bY would be "qualitatively" the same as that of X and Y.

To see the "qualitative" relation between X and Y, we normalize it by dividing it by the covariance by the sum of the standard deviations of X and Y. The normalized covariance is called the ***correlation coefficient*** of X and Y.

Definition of the correlation coefficient

Definition (Correlation coefficient)

Let X and Y be random variables. The ***correlation coefficient*** $\mathrm{corr}[X,Y]$ between X and Y is given by

$$\mathrm{corr}[X,Y] := \frac{\mathrm{Cov}[X,Y]}{\sigma[X]\sigma[Y]}. \tag{20}$$

As expected, for positive real numbers a and b, we have that

$$\mathrm{corr}[aX,bY] = \mathrm{corr}[X,Y]. \tag{21}$$

Correlation $\neq$ Causality

If two random variables X and Y have a correlation, i.e., $\mathrm{corr}[X, Y] \neq 0$, you might expect that X is the cause of Y.

However, there are many possibilities behind the correlation, e.g.,

1. X is a cause of Y.
2. Y is a cause of X.
3. There exists a random variable Z that causes the both X and Y.
4. (When we estimate the correlation coefficient) There is no relation between X and Y but our estimation of the correlation coefficient is non-zero by estimation errors.

Hence, we cannot conclude that X is a cause of Y just by $\mathrm{corr}[X, Y] \neq 0$.

Continuous random variables in real AI applications

- Prices of goods, stocks, etc. (Economic data)
- RGB values of each pixel in an image.
- The intensity of an acoustic signal at each time frame.
- Internal states of neural networks.

A random variable may not have a PMF.

Consider a simple random variable uniformly distributed in $[0, 1]$. Here $\Pr(0 \leq X \leq 1) = 1$.

This random variable have nowhere probability mass, i.e., $\Pr(X = x) = 0.$ for any $X \in \mathbb{R}$.

Proof.

Since its support is $[0, 1]$, it is trivial that $\Pr(X = x) = 0$ for $x \neq [0, 1]$. For $x \in [0, 1]$, assume, for the sake of contradiction, that $\Pr(X = x) = \epsilon$, where $\epsilon > 0$. From its uniformity, if $\Pr(X = x) = \epsilon$ holds for one value $x \in [0, 1]$, then it holds for all $x \in [0, 1]$. Hence, if $A \subset [0, 1]$ and A has at least N elements, $\Pr(X \in A) \leq N\epsilon$. However, there are a infinite number of real numbers in $[0, 1]$, so $\Pr(X \in [0, 1])$ is infinity. It contradicts $\Pr(X \in [0, 1]) = 1$. □

A random variable may not have a PMF.

Consider a simple random variable uniformly distributed in $[0, 1]$. Here $\Pr(0 \leq X \leq 1) = 1$.

This random variable have nowhere probability mass, i.e., $\Pr(X = x) = 0.$ for any $X \in \mathbb{R}$.

Random variables whose support is a section in the real line have the same problem. Hence, we need another way to represent a random variable.

Cumulative distribution function (CDF)

Any random variable have a ***cumulative distribution function (CDF)*** defined as follows.

Definition

Let X be a random variable. The ***cumulative distribution function (CDF)*** $F_X : \mathbb{R} \to [0, 1]$ of X is defined by

$$F_X(x) := \Pr(X \leq x). \tag{22}$$

CDF of rolling an ideal dice

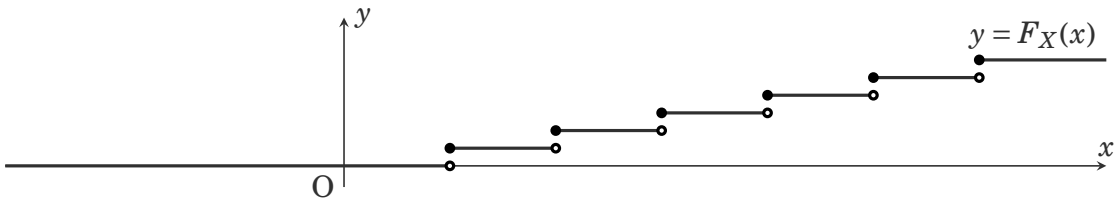
Suppose that we roll an ideal six-sided dice. The PMF is given as follows.

x	1	2	3	4	5	6
$P_X(x) := \Pr(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

The PMF of rolling an ideal six-sided dice

CDF of rolling an ideal dice

Suppose that we roll an ideal six-sided dice. The CDF is given as follows.

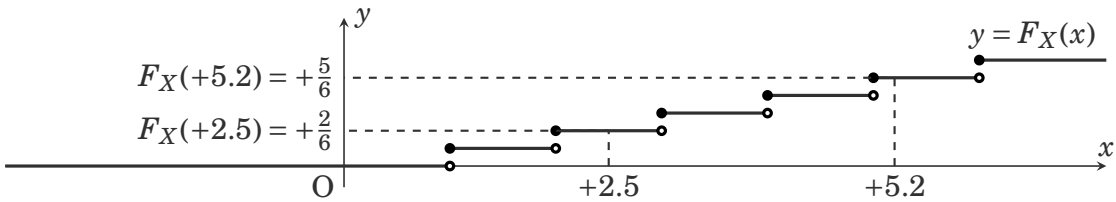


x	$(-\infty, 1)$	$[1,2)$	$[2,3)$	$[3,4)$	$[4,5)$	$[5,6)$	$[6,+\infty)$
$F_X(x) := \Pr(X = x)$	0	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1

The CDF of rolling an ideal six-sided dice

CDF of rolling an ideal dice

Suppose that we roll an ideal six-sided dice. The CDF is given as follows.

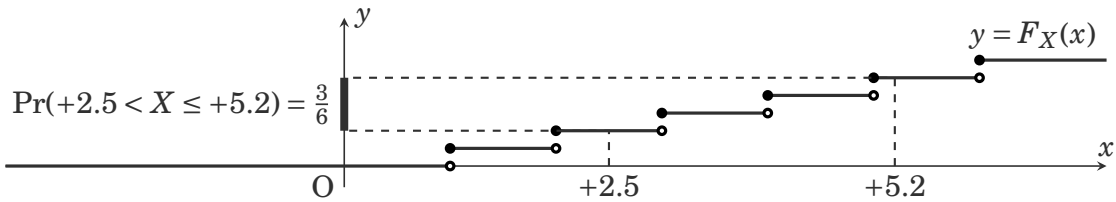


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned} \Pr(+2.5 < X \leq +5.2) &= \Pr(X \leq +5.2) - \Pr(X \leq +2.5) \\ &= F_X(+5.2) - F_X(+2.5) \\ &= \frac{5}{6} - \frac{2}{6} = \frac{3}{6}. \end{aligned} \tag{23}$$

CDF of rolling an ideal dice

Suppose that we roll an ideal six-sided dice. The CDF is given as follows.

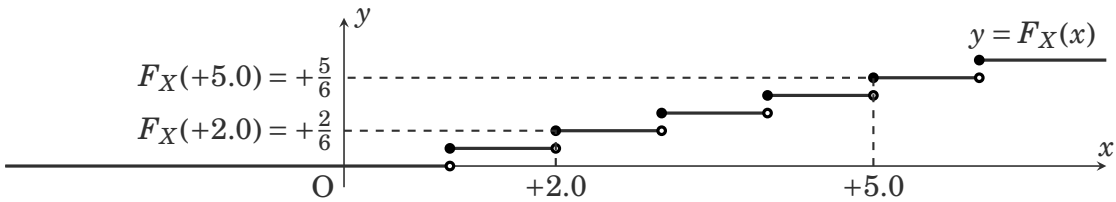


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned}\Pr(+2.5 < X \leq +5.2) &= \Pr(X \leq +5.2) - \Pr(X \leq +2.5) \\ &= F_X(+5.2) - F_X(+2.5) \\ &= \frac{5}{6} - \frac{2}{6} = \frac{3}{6}.\end{aligned} \tag{23}$$

CDF of rolling an ideal dice

Suppose that we roll an ideal six-sided dice. The CDF is given as follows.

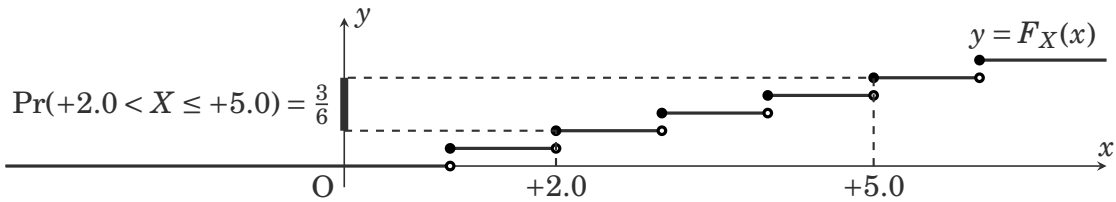


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned} \Pr(+2.0 < X \leq +5.0) &= \Pr(X \leq +5.0) - \Pr(X \leq +2.0) \\ &= F_X(+5.0) - F_X(+2.0) \\ &= \frac{5}{6} - \frac{2}{6} = \frac{3}{6}. \end{aligned} \tag{23}$$

CDF of rolling an ideal dice

Suppose that we roll an ideal six-sided dice. The CDF is given as follows.

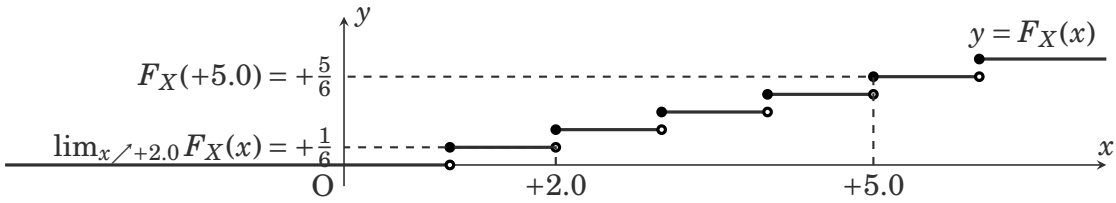


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned} \Pr(+2.0 < X \leq +5.0) &= \Pr(X \leq +5.0) - \Pr(X \leq +2.0) \\ &= F_X(+5.0) - F_X(+2.0) \\ &= \frac{5}{6} - \frac{2}{6} = \frac{3}{6}. \end{aligned} \tag{23}$$

CDF of rolling an ideal dice

Suppose that we roll an ideal six-sided dice. The CDF is given as follows.

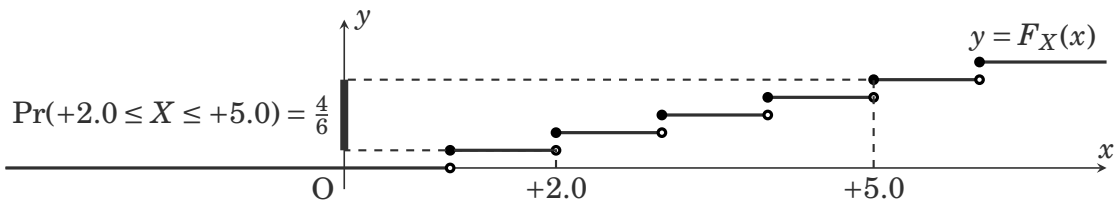


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned} \Pr(+2.0 \le X \le +5.0) &= \Pr(X \le +5.0) - \Pr(X < +2.0) \\ &= F_X(+5.0) - \lim_{x \nearrow +2.0} F_X(x) \\ &= \frac{5}{6} - \frac{1}{6} = \frac{4}{6}. \end{aligned} \tag{23}$$

CDF of rolling an ideal dice

Suppose that we roll an ideal six-sided dice. The CDF is given as follows.

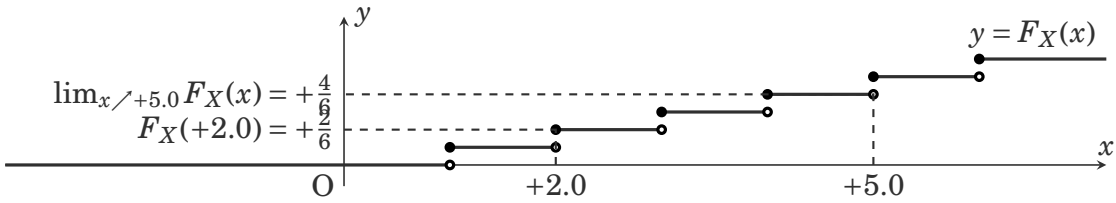


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned}\Pr(+2.0 \le X \le +5.0) &= \Pr(X \le +5.0) - \Pr(X < +2.0) \\ &= F_X(+5.0) - \lim_{x \nearrow +2.0} F_X(x) \\ &= \frac{5}{6} - \frac{1}{6} = \frac{4}{6}.\end{aligned} \tag{23}$$

CDF of rolling an ideal dice

Suppose that we roll an ideal six-sided dice. The CDF is given as follows.

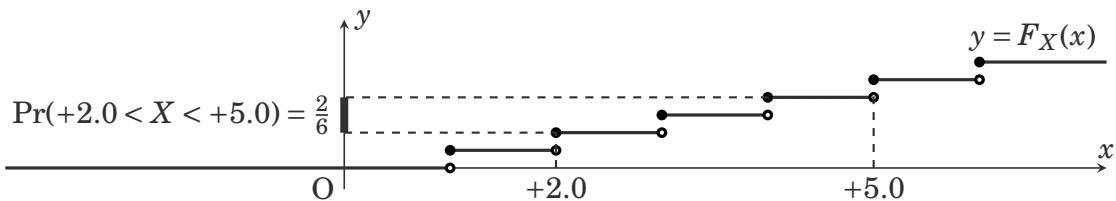


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned} \Pr(+2.0 < X < +5.0) &= \Pr(X < +5.0) - \Pr(X \leq +2.0) \\ &= \lim_{x \nearrow +5.0} F_X(x) - F_X(+2.0) \\ &= \frac{4}{6} - \frac{2}{6} = \frac{2}{6}. \end{aligned} \tag{23}$$

CDF of rolling an ideal dice

Suppose that we roll an ideal six-sided dice. The CDF is given as follows.

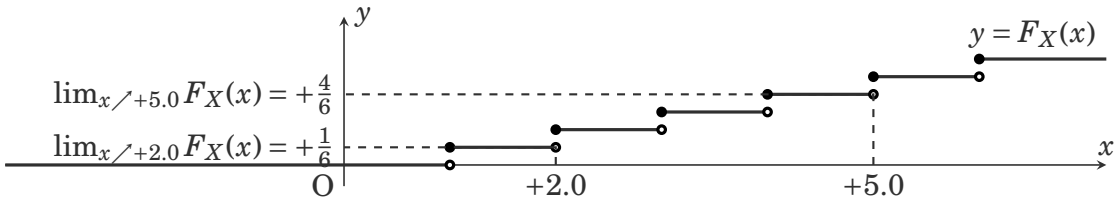


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned}\Pr(+2.0 < X < +5.0) &= \Pr(X < +5.0) - \Pr(X \leq +2.0) \\ &= \lim_{x \nearrow +5.0} F_X(x) - F_X(+2.0) \\ &= \frac{4}{6} - \frac{2}{6} = \frac{2}{6}.\end{aligned} \tag{23}$$

CDF of rolling an ideal dice

Suppose that we roll an ideal six-sided dice. The CDF is given as follows.

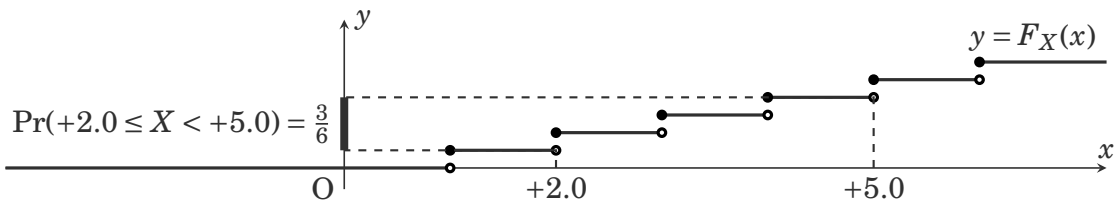


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned} \Pr(+2.0 \le X < +5.0) &= \Pr(X < +5.0) - \Pr(X < +2.0) \\ &= \lim_{x \nearrow +5.0} F_X(x) - \lim_{x \nearrow +2.0} F_X(x) \\ &= \frac{4}{6} - \frac{1}{6} = \frac{3}{6}. \end{aligned} \tag{23}$$

CDF of rolling an ideal dice

Suppose that we roll an ideal six-sided dice. The CDF is given as follows.

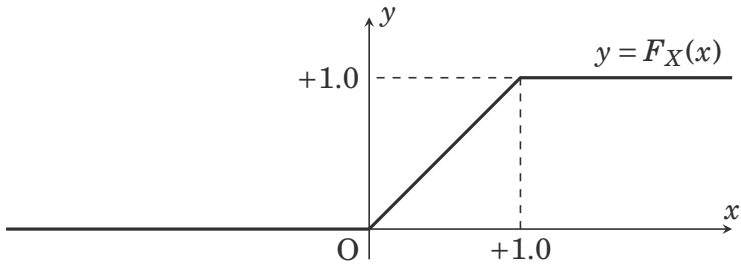


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned}\Pr(+2.0 \le X < +5.0) &= \Pr(X < +5.0) - \Pr(X < +2.0) \\ &= \lim_{x \nearrow +5.0} F_X(x) - \lim_{x \nearrow +2.0} F_X(x) \\ &= \frac{4}{6} - \frac{1}{6} = \frac{3}{6}.\end{aligned} \tag{23}$$

CDF example

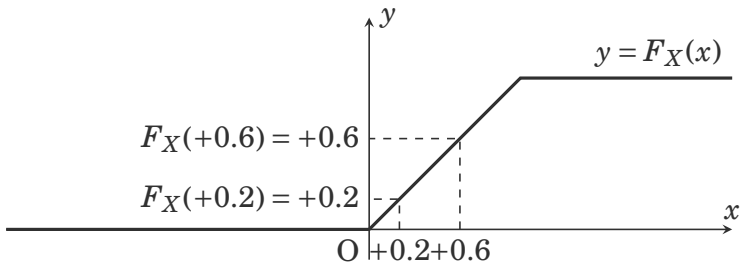
The CDF of a random variable X uniformly distributed in $[0, 1]$ is:



$$F_X(x) = \begin{cases} 0 & \text{if } x \le 0, \\ x & \text{if } 0 \le x \le 1, \\ 1 & \text{if } x \ge 1. \end{cases} \tag{24}$$

CDF example

The CDF of a random variable X uniformly distributed in $[0, 1]$ is:

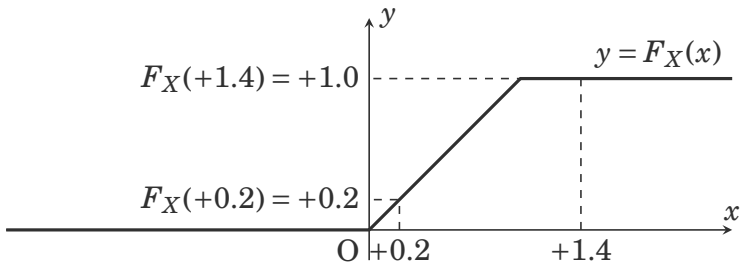


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned} \Pr(0.2 < X \leq 0.6) &= \Pr(X \leq 0.6) - \Pr(X \leq 0.2) \\ &= F_X(0.6) - F_X(0.2) \\ &= 0.6 - 0.2 = 0.4. \end{aligned} \tag{24}$$

CDF example

The CDF of a random variable X uniformly distributed in $[0, 1]$ is:

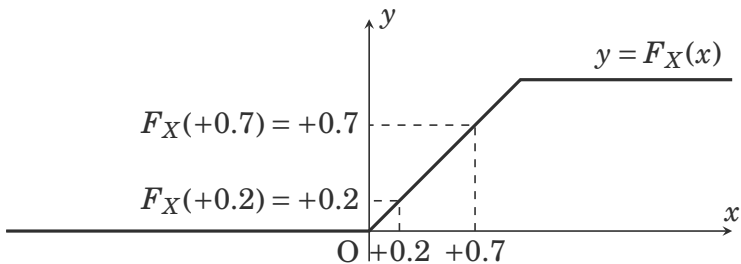


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned} \Pr(0.2 \le X \le 0.7) &= \Pr(X \le 0.7) - \lim_{x \nearrow 0.2} \Pr(x) \\ &= F_X(0.7) - \lim_{x \nearrow 0.2} F_X(x) \\ &= 0.7 - 0.2 = 0.5. \end{aligned} \tag{24}$$

CDF example

The CDF of a random variable X uniformly distributed in $[0, 1]$ is:



Using the CDF, we can calculate the probability of various events. For example,

Properties of CDF

For any random variable X, its CDF F_X satisfies

- $\lim_{x \to -\infty} F_X(x) = 0$.
- $\lim_{x \to +\infty} F_X(x) = 1$.
- The CDF is everywhere right-continuous, i.e., $\lim_{x \nearrow x_0} F_X(x) = F_X(x_0)$ for all $x_0 \in \mathbb{R}$.
- The CDF has its left-limit $\lim_{x \searrow x_0} F_X(x)$ for all $x_0 \in \mathbb{R}$.

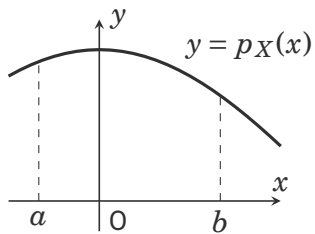
Why we are not satisfied with a CDF?

The CDF is not intuitive. At one glance, we do not know around which value the random variable tends to take a value.

Probability density function (PDF): an infinitely precise histogram

Suppose that $a \leq b$.

Given a probability density function p_X, the probability of the random variable X taking a value between a and b is given by the area bounded by the graph of $y = p(x)$ and $y = 0$ between $x = a$ and $x = b$.

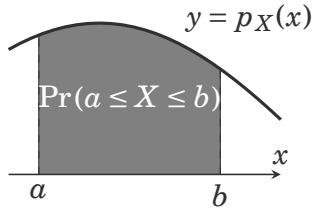


If the probability density function (PDF) of a random variable is given, the probability $\Pr(a \leq X \leq b)$ is given by the area under the PDF in the domain $[a, b]$.

Probability density function (PDF): an infinitely precise histogram

Suppose that $a \leq b$.

Given a probability density function p_X, the probability of the random variable X taking a value between a and b is given by the area bounded by the graph of $y = p(x)$ and $y = 0$ between $x = a$ and $x = b$.



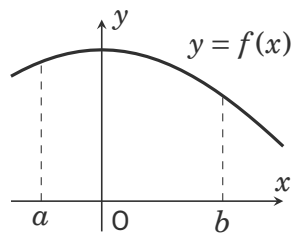
If the probability density function (PDF) of a random variable is given, the probability $\Pr(a \leq X \leq b)$ is given by the area under the PDF in the domain $[a, b]$.

Definite Integral

Suppose that $a \leq b$.

The (signed) area bounded by the graph of $y = f(x)$ and $y = 0$ between $x = a$ and $x = b$ is called the ***definite integral*** of f between a and b, which is denoted by $\int_a^b f(x)\,\mathrm{d}x$.

We also define $\int_b^a f(x)\,\mathrm{d}x := -\int_a^b f(x)\,\mathrm{d}x$.



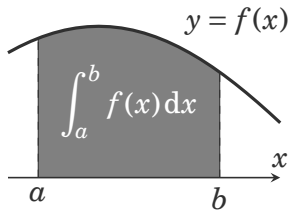
The definite integral is the area bounded by the graph of the function.

Definite Integral

Suppose that $a \leq b$.

The (signed) area bounded by the graph of $y = f(x)$ and $y = 0$ between $x = a$ and $x = b$ is called the ***definite integral*** of f between a and b, which is denoted by $\int_a^b f(x)\,\mathrm{d}x$.

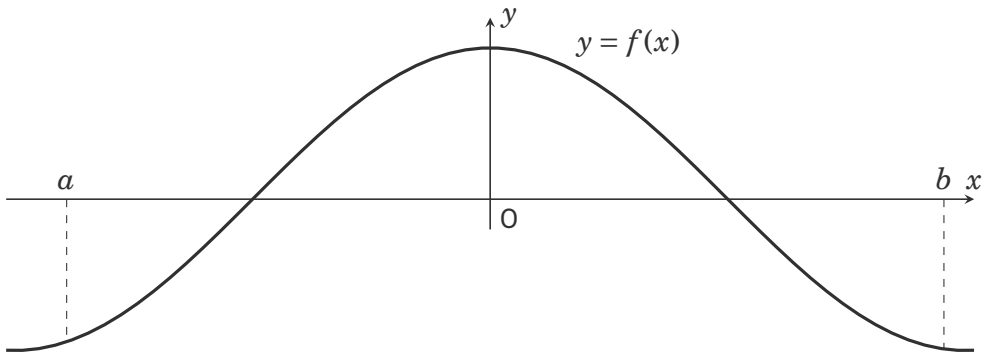
We also define $\int_b^a f(x)\,\mathrm{d}x := -\int_a^b f(x)\,\mathrm{d}x$.



The definite integral is the area bounded by the graph of the function.

Definite Integral: When the function takes negative values

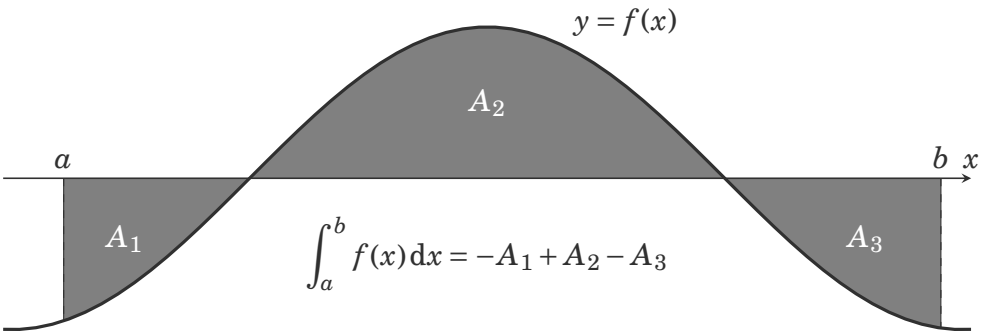
Areas bounded by the graph taking negative values are counted as negative values.



Areas bounded by the graph taking negative values are counted as negative values.

Definite Integral: When the function takes negative values

Areas bounded by the graph taking negative values are counted as negative values.



Areas bounded by the graph taking negative values are counted as negative values.

Improper integral

A random variable may take all the real values. Hence, we often consider the area bounded by a function's graph in domains like $(-\infty, +\infty)$.

The area bounded by a graph in an infinite size section is called an ***improper integral***, defined as follows.

- $\int_a^{+\infty} f(x)\,\mathrm{d}x := \lim_{b \to +\infty} \int_a^b f(x)\,\mathrm{d}x.$
- $\int_{-\infty}^b f(x)\,\mathrm{d}x := \lim_{a \to -\infty} \int_a^b f(x)\,\mathrm{d}x.$
- $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x := \int_{-\infty}^c f(x)\,\mathrm{d}x + \int_c^{+\infty} f(x)\,\mathrm{d}x$, where c is an arbitrary real value[5].

[5] The selection of the value c does not change the result.

Properties of the definite integral

Let a, b, c be real numbers and f and g be functions of a real value.

- $\int_b^a f(x)\,\mathrm{d}x := -\int_a^b f(x)\,\mathrm{d}x.$
- $\int_a^a f(x)\,\mathrm{d}x = 0.$
- $\int_a^b [f(x) + g(x)]\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x + \int_a^b g(x)\,\mathrm{d}x.$
- $\int_a^b cf(x)\,\mathrm{d}x = c\int_a^b f(x)\,\mathrm{d}x.$
- $\int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x.$

Other applications of definite integral

Consider a car whose velocity at time t is given by $v(t)$. Let the position of the car at time 0 be 0 then the position $x(t)$ at time t is given by

$$x(t) = \int_0^t v(t)\mathrm{d}t. \tag{24}$$

Expectation (mean) of a continuous random variable

If the probability density function of a random variable X is given by p, then the expectation of X is given by

$$\int_{-\infty}^{+\infty} x p(x) \mathrm{d}x. \tag{25}$$

Cf.) The expectation of a discrete random variable X is given by

$$\sum_{x \in \mathcal{X}} x P(x), \tag{26}$$

where P is the probability mass function.

Expectation of the value of a function

If the probability density function of a random variable X is given by p, then the expectation of $f(X)$ is given by

$$\int_{-\infty}^{+\infty} f(x)p(x)\mathrm{d}x. \tag{27}$$

Cf.) The expectation of a discrete random variable $f(X)$ is given by

$$\sum_{x\in\mathbb{X}} f(x)P(x), \tag{28}$$

where P is the probability mass function.

The expectation does not always exist

If the PDF of a random variable X is given by

$$p(x) = \frac{1}{1+x^2}, \tag{29}$$

then X does not have its expectation. Indeed, the improper integral

$$\int_{-\infty}^{+\infty} xp(x)\mathrm{d}x := \lim_{a\to-\infty}\int_a^c xp(x)\mathrm{d}x + \lim_{b\to+\infty}\int_c^b xp(x)\mathrm{d}x \tag{30}$$

diverges (both the first and second terms in the RHS diverge).

Variance and standard deviation of a continuous random variable

The variance of a random variable is given by the expectation of the square deviation; that is

$$\int_{-\infty}^{+\infty} (x-m)^2 p(x) \mathrm{d}x. \tag{31}$$

The standard deviation is given by the square root of the variance.

Calculating definite integrals

- Numerical integration
- Calculating analytically as the inverse operation of differentiation

Integral is the “inverse” of differentiation

Definition (Primitive function)

Let a and b be real numbers such that $a < b$ and $f : [a, b] \to \mathbb{R}$. If $F : [a, b] \to \mathbb{R}$ satisfies $F' = f$, i.e., $\frac{\mathrm{d}}{\mathrm{d}x} F(x) = f(x)$ for all $x \in [a, b]$, then F is called a ***primitive function*** or an ***antiderivative function*** of f.

Theorem (The fundamental theorem of calculus (FTC))

Let a and b be real numbers such that $a < b$ and $f : [a, b] \to \mathbb{R}$ be integrable. Suppose that there exists a function $F : [a, b] \to \mathbb{R}$, then we have that

$$\int_a^x f(t)\mathrm{d}t = F(x) - F(a). \tag{32}$$

According to the FTC, we can **calculate an integral using a primitive function**!

Calculating the definite integral

To calculate the definite integral

$$\int_a^b f(x)\mathrm{d}x, \tag{33}$$

we will

1. Find a primitive (antiderivative) function $F : [a, b] \to \mathbb{R}$, which satisfies $F' = f$.
2. Find the value of $F(b) - F(a)$.

A primitive function is not unique but unique up to constant.

If a function $F_1 : [a, b] \to \mathbb{R}$ is a primitive function of $f : [a, b] \to \mathbb{R}$, then $F_2 : [a, b] \to \mathbb{R}$ defined by $F_2(x) = F_1(x) + C$ is also a primitive function, where $C \in \mathbb{R}$ is a constant.

Example

Both $F_1(x) = \frac{1}{2}x^2$ and $F_2(x) = \frac{1}{2}x^2 + 5$ are primitive functions of $f(x) = x$.

On the other hand, if both F_1 and F_2 are primitive functions of f, then the difference between F_1 and F_2 is a constant function.

In this sense, we say that the primitive function is unique up to a constant, and to denote all the primitive functions, we write like $F(x) + C$ using a primitive function $F(x)$. Here, the constant C is called the constant of integration.

Example

The primitive function of $f(x) = x$ is $\frac{1}{2}x^2 + C$.

Primitive function of a simple function might be complicated.

The primitive function of $\exp\left(-x^2\right)$ is known to be impossible to represent as an elementary function, that is, a function defined as taking sums, products, roots and compositions of finitely many polynomial, rational, trigonometric, and exponential functions (though the original function $\exp\left(-x^2\right)$ itself is an elementary function).

From the computer science viewpoint, some non-elementary primitive functions might be implemented by some libraries if they are famous. If they are not implemented, you might need to calculate the definite integral using a numerical method.

Multi integral

Let $D = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_m, b_m]$ be a m-dimensional hyper-cube. Let $f : D \to \mathbb{R}$ be a function of a m-dimensional variable. Similar to one-dimensional function cases, we call the set of points

$$\{(x_1, x_2, \ldots, x_m, f(x_1, x_2, \ldots, x_m)) | (x_1, x_2, \ldots, x_m) \in D\} \tag{34}$$

the ***graph*** of a function f. The (signed) volume in the domain D bounded by the graph of $y = f(\boldsymbol{x})$ and $y = 0$ is called the ***multi integral*** of f on D, denoted by $\int_D f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$.

For general domain $A \subset D$, we define the multi integral of f on D by

$$\int_A f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} := \int_D 1_A(\boldsymbol{x}) f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{35}$$

Joint PDF

Let $X_1, X_2, \ldots, X_m$ be random variables. If $p_{X_1,X_2,\ldots,X_m} : \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ satisfies

$$\Pr((X_1, X_2, \ldots, X_m) \in A) = \int_A p_{X_1,X_2,\ldots,X_m}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \tag{36}$$

then the function $p_{X_1,X_2,\ldots,X_m}$ is called the ***joint probability density function (joint PDF)*** of $X_1, X_2, \ldots, X_m$.

Marginal PDF (bivariable cases)

Suppose that the joint PDF $p_{X,Y}$ is given. The ***marginal probability density functions (marginal PDFs)*** p_X and p_Y are given by

$$
\begin{aligned}
p_X(x) &= \int_{\infty}^{\infty} p_{X,Y}(x,y)\mathrm{d}y, \\
p_Y(y) &= \int_{\infty}^{\infty} p_{X,Y}(x,y)\mathrm{d}x.
\end{aligned}
\tag{37}
$$

Conditional PDF (bivariate cases)

The ***conditional probability distribution function (conditional PDF)*** is defined by

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}, \quad p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}. \tag{38}$$

Calculating the expectation of a function from joint PDF

Let $X_1, X_2, \ldots, X_m$ be random variables and $p_{X_1,X_2,\ldots,X_m}$ be the joint PDF. Let $f : \mathbb{R}^m \to \mathbb{R}$ be a function. The expectation of the random variable $f(X_1, X_2, \ldots, X_m)$ is given by

$$\int_{\mathbb{R}^m} f(\boldsymbol{x}) p_{X_1,X_2,\ldots,X_m}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{39}$$

Covariance

Let X and Y are random variables and μ_X and μ_Y be the expectation of X and Y, respectively. Suppose that $p_{X,Y}$ is a joint PDF of X and Y. Then, the covariance $\mathrm{Cov}(X,Y)$ is given by

$$\mathrm{Cov}(X,Y) = \int_{\mathbb{R}^2} \big(x-\mu_X\big)\big(y-\mu_Y\big) p_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y \tag{40}$$

Calculating multi integral by iterated integral

We can calculate a multi-integral by an ***iterated integral***.

Theorem

Under some loose conditions[6]*, we have that*

$$\begin{aligned}
&\iint_A p(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} 1_A(x,y)p(x,y)\mathrm{d}x\right]\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} 1_A(x,y)p(x,y)\mathrm{d}y\right]\mathrm{d}x.
\end{aligned} \tag{41}$$

[6] We refer the readers wanting to know the exact conditions to the Fubini-Tonelli theorem.

Sample and sample statistics

In real applications, we **rarely know the true distribution**, behind the data.

On the other hand, we often **have many data points** that we can assume follow the same distribution (often independently). Such a series of data points is called ***sample*** of the distribution.

Statistics, data science, machine learning, etc., aim to **extract information about the true distribution from available data points. Sample statistics are the basis of those pieces of technology**.

Terminology: population and sample

In the context of statistics,

- The true distribution is often called the ***population***.
- A series of data points that we can assume follow the same distribution is called ***sample***. If it has many data points, we say that the sample is large, and if it has few data, we say that the sample is small.

Summary statistics and sample statistics

- ***Summary statistics*** aims to describe characteristics of a (known or true) distribution by a few values.
- ***Sample statistics*** aims to estimate some information about the true distribution from finite sample data.

We only have finite data points, so sample statistics are practically necessary to handle probability in real applications.

Sample mean

One principal summary statistic is the expectation.

For data points $X_1, X_2, \ldots, X_m$, we can easily calculate the ***sample mean***

$$m_m = \frac{1}{m}(X_1 + X_2 + \cdots + X_m), \tag{42}$$

the mean of the data points.

If we can assume that those data points are the values of random variables following the same distribution with a true mean μ, we expect m to approximate the true mean μ, which is unknown.

Is it correct? The answer is YES, according to the **law of large numbers**.

Law of large numbers

Theorem ((Strong) law of large numbers)

Let $X_1, X_2, \dots$ be an infinite sequence of independently and identically distributed (i.i.d.) random variables and assume that the mean of the distribution is $\mu \in \mathbb{R}$.
Let $\overline{X}_m$ be the sample mean

$$\overline{X}_m := \frac{1}{m}(X_1 + X_2 + \cdots + X_m). \tag{43}$$

Then $\overline{X}_m$ converges to μ in probability 1.

Thus, the sample mean tells us some information about the unknown true distribution!

How the sample mean behaves?

The sample mean converges to the expectation. Now,

- How close to the expectation will the sample mean get as we increase the data points?
- What does the distribution of the sample mean look like?

The answer is

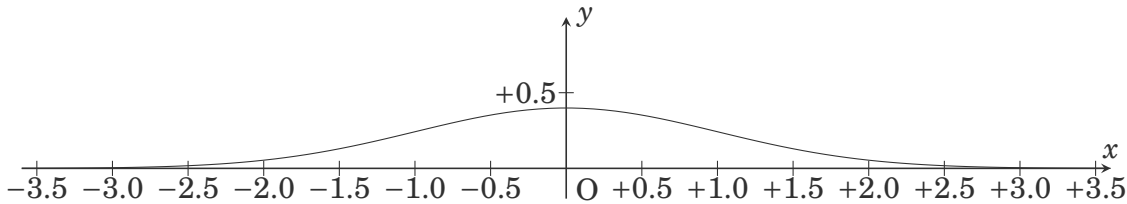
- The difference between the sample mean and the true expectation is proportional to the standard deviation σ of the true distribution and $\frac{1}{\sqrt{m}}$,
- With appropriate scaling, the distribution of the sample mean converges to a ***normal distribution (Gaussian distribution)***,

according to the ***central limit theorem***.

What is the normal distribution?

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{44}$$



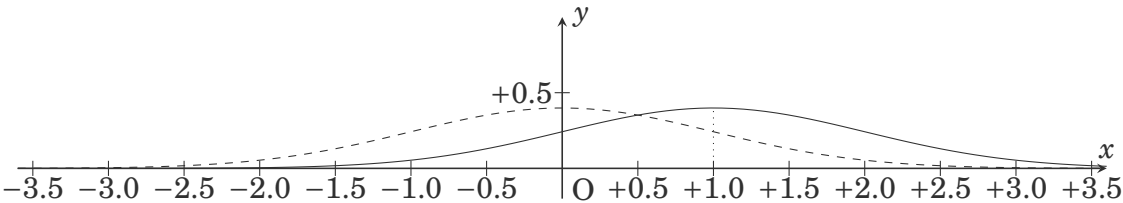
Normal distributions' PDF ($\mu = 0, \sigma = 1$).

The mean, the variance, and the standard deviation are μ, σ^2, and $\sigma := \sqrt{\sigma^2}$, respectively.

What is the normal distribution?

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{44}$$



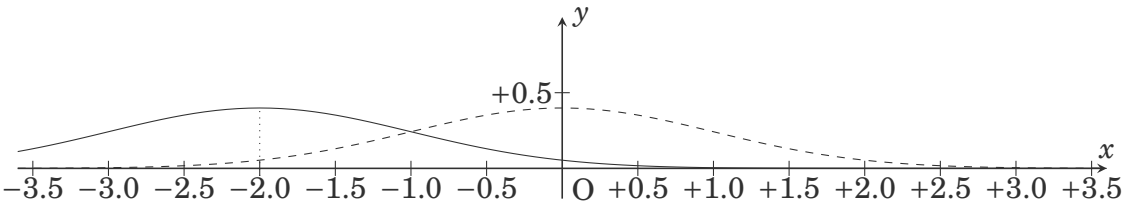
Normal distributions' PDF (Solid: $\mu = 1, \sigma = 1$, Dashed: $\mu = 0, \sigma = 1$).

The mean, the variance, and the standard deviation are μ, σ^2, and $\sigma := \sqrt{\sigma^2}$, respectively.

What is the normal distribution?

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{44}$$



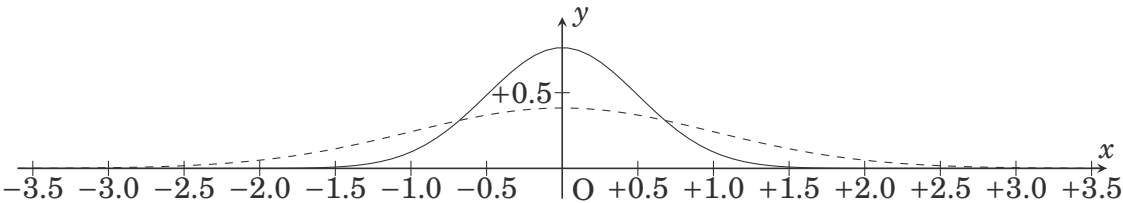
Normal distributions' PDF (Solid: $\mu = -2, \sigma = 1$, Dashed: $\mu = 0, \sigma = 1$).

The mean, the variance, and the standard deviation are μ, σ^2, and $\sigma := \sqrt{\sigma^2}$, respectively.

What is the normal distribution?

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{44}$$



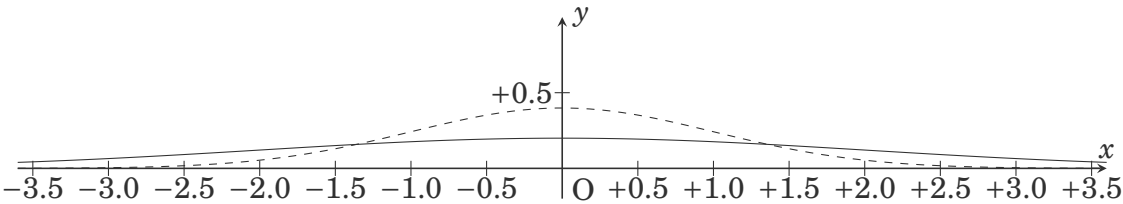
Normal distributions' PDF (Solid: $\mu = 0, \sigma = 0.5$, Dashed: $\mu = 0, \sigma = 1$).

The mean, the variance, and the standard deviation are μ, σ^2, and $\sigma := \sqrt{\sigma^2}$, respectively.

What is the normal distribution?

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{44}$$



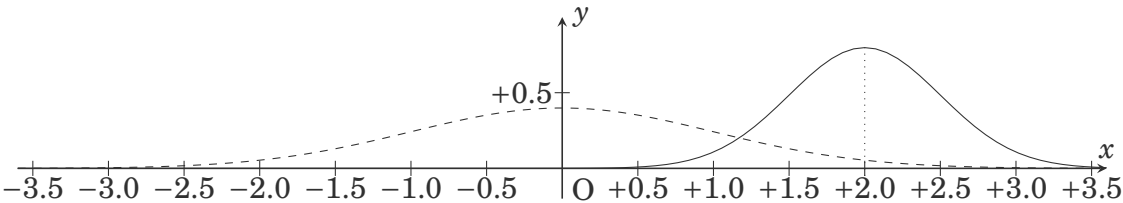
Normal distributions' PDF (Solid: $\mu = 0, \sigma = 2.0$, Dashed: $\mu = 0, \sigma = 1$).

The mean, the variance, and the standard deviation are μ, σ^2, and $\sigma := \sqrt{\sigma^2}$, respectively.

What is the normal distribution?

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{44}$$



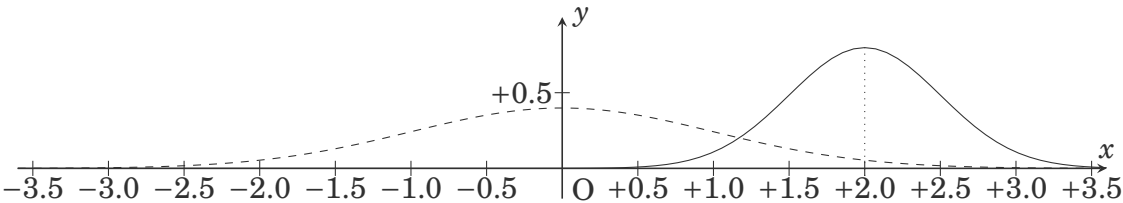
Normal distributions' PDF (Solid: $\mu = 2, \sigma = 0.5$, Dashed: $\mu = 0, \sigma = 1$).

The mean, the variance, and the standard deviation are μ, σ^2, and $\sigma := \sqrt{\sigma^2}$, respectively.

What is the normal distribution?

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{44}$$



Normal distributions' PDF (Solid: $\mu = 2, \sigma = 0.5$, Dashed: $\mu = 0, \sigma = 1$).

The PDF is symmetric about $x = \mu$ and it is dense around $x = \mu$.

Central limit theorem (CLT)

Theorem (Central limit theorem (CLT))

Let $X_1, X_2, \ldots$ be an infinite sequence of independently and identically distributed (i.i.d.) random variables and assume that the mean and variance of the distribution are $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_{\geq 0}$, respectively.
Let $\overline{X}_m$ be the sample mean

$$\overline{X}_m := \frac{1}{m}(X_1 + X_2 + \cdots + X_m). \tag{45}$$

Then the CDF of $\sqrt{m}\frac{\overline{X}_m - \mu}{\sigma}$ converges to that of the standard normal distribution at any point in $\mathbb{R}$.

The implications of the CLT

- The error $\overline{X}_m - \mu$ in estimating the true mean μ is almost proportional to $\frac{1}{\sqrt{m}}$. In particular, the more data points, the more accurate the estimate is.
- The sum of sufficiently many independent random variables approximately follows a normal distribution. In particular, various types of random variables decomposable to many independent factors follow a normal distribution. This is why **the normal distribution appears everywhere in the real world**.

Estimation of a distribution

We have estimated the expectation only. In real applications, we might want to estimate the distribution itself. However, if the support of the distribution is infinite, it is not practical to determine a PMF or PDF from finite data points with no assumptions.

We often assume that the distribution is in a parametric model, which is a set of distributions parametrized by a few values.

Parametric model

Definition (A parametric model)

- ***A discrete parametric model*** on support $\mathscr{X} \subset \mathbb{R}^n$ is a pair of a parameter set $\Theta \subset \mathbb{R}^k$ and a parametrized PMF $P : \mathscr{X} \times \Theta \to [0, 1]$ such that $P(\boldsymbol{x}; \boldsymbol{\theta})$ is a PMF on $\mathscr{X}$ as a function of $\boldsymbol{x}$ for all $\boldsymbol{\theta} \in \Theta$.
- ***A continuous parametric model*** on support $\mathbb{R}^n$ is a pair of a parameter set $\Theta \subset \mathbb{R}^k$ and a parametrized PDF $p : \mathbb{R}^n \times \Theta \to \mathbb{R}_{\geq 0}$ such that $p(\boldsymbol{x}; \boldsymbol{\theta})$ is a PDF on $\mathbb{R}^n$ as a function of $\boldsymbol{x}$ for all $\boldsymbol{\theta} \in \Theta$.

Here, the nonnegative integer k is the dimension of the parameter.

When we have a parametric model, estimating a parameter corresponds to estimating a distribution.

Likelihood

To determine a parameter of a parametric model from data points, we quantify how "likely" the distribution indicated by a parameter is correct.

When we have a PMF or PDF of a distribution, we simply define the value of the PMF or PDF of the data points as the ***likelihood*** of the distribution.

Definition (Likelihood of a discrete parametric model)

Let $P(\cdot;\cdot)$ be a discrete parametric model with a parameter set Θ and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ be values of data points.
Then the ***likelihood*** of $P(\cdot;\boldsymbol{\theta})$ (or often called the likelihood of the parameter $\boldsymbol{\theta}$) is defined as the following product.

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{46}$$

Likelihood

To determine a parameter of a parametric model from data points, we quantify how "likely" the distribution indicated by a parameter is correct.

When we have a PMF or PDF of a distribution, we simply define the value of the PMF or PDF of the data points as the ***likelihood*** of the distribution.

Definition (Likelihood of a continuous parametric model)

Let $p(\cdot;\cdot)$ be a continuous parametric model with a parameter set Θ and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ be values of data points.
Then the ***likelihood*** of $p(\cdot;\boldsymbol{\theta})$ (or often called the likelihood of the parameter $\boldsymbol{\theta}$) is defined as the following product.

$$p(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot p(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot p(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{46}$$

Probability and likelihood

The value of the product

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}) \tag{47}$$

can be interpreted as either

- the probability of the random variable sequence taking the value sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$, i.e., a function of a value sequence, or
- the likelihood of the distribution determined by the parameter $\boldsymbol{\theta}$, i.e., a function of a distribution (or parameter).

In other words, the above product is the probability (or the probability density for continuous distribution case) if we interpret it as a function of a value sequence, and the likelihood if we interpret it as a function of a distribution (or a parameter).

Maximum likelihood estimator

Once we define the likelihood of a distribution, all we need to do is to find a parameter that maximizes the likelihood.

The parameter vector that maximizes the likelihood is called the ***maximum likelihood estimator (MLE)***.

Definition (Maximum likelihood estimator)

Let $P(\cdot;\cdot)$ be a discrete parametric model with a parameter set Θ and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ be values of data points.
The parameter vector $\boldsymbol{\theta}$ is called a maximum likelihood estimator (MLE) if it maximizes the likelihood

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{48}$$

If there is a unique MLE, we often denote it by $\hat{\boldsymbol{\theta}}$.

MLE maximizes the score and minimizes the negative log likelihood

For a parameter vector $\boldsymbol{\theta}$, the following is equivalent[7].

- The parameter vector $\boldsymbol{\theta}$ maximizes the likelihood function

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{49}$$

- The parameter vector $\boldsymbol{\theta}$ maximizes the ***log likelihood*** function

$$\log P(\boldsymbol{x}_1;\boldsymbol{\theta}) + \log P(\boldsymbol{x}_2;\boldsymbol{\theta}) + \cdots + \log P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{50}$$

- The parameter vector $\boldsymbol{\theta}$ minimizes the ***negative log likelihood*** function

$$-\log P(\boldsymbol{x}_1;\boldsymbol{\theta}) - \log P(\boldsymbol{x}_2;\boldsymbol{\theta}) - \cdots - \log P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{51}$$

[7] It follows since $\log$ is an increasing function. It holds regardless of the base of the logarithm.

MLE of normal distribution model is minimizing squared error.

Let $p(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Then, the negative (natural) log likelihood of the data sequence is given by

$$\begin{aligned} &\log\left(2\pi\sigma^2\right) + \frac{\left(x_1-\mu\right)^2}{2\sigma^2} + \log\left(2\pi\sigma^2\right) + \frac{\left(x_2-\mu\right)^2}{2\sigma^2} + \cdots + \log\left(2\pi\sigma^2\right) + \frac{\left(x_m-\mu\right)^2}{2\sigma^2} \\ &= m\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left[\left(x_1-\mu\right)^2 + \left(x_2-\mu\right)^2 + \cdots + \left(x_m-\mu\right)^2\right]. \end{aligned} \tag{52}$$

MLE of normal distribution model is minimizing squared error.

Let $p(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Then, the negative (natural) log likelihood of the data sequence is given by

$$\begin{aligned} &\log\left(2\pi\sigma^2\right) + \frac{\left(x_1-\mu\right)^2}{2\sigma^2} + \log\left(2\pi\sigma^2\right) + \frac{\left(x_2-\mu\right)^2}{2\sigma^2} + \cdots + \log\left(2\pi\sigma^2\right) + \frac{\left(x_m-\mu\right)^2}{2\sigma^2} \\ &= m\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left[\left(x_1-\mu\right)^2 + \left(x_2-\mu\right)^2 + \cdots + \left(x_m-\mu\right)^2\right]. \end{aligned} \tag{52}$$

When we minimize the above with respect to μ, we can ignore the gray parts.

In this sense, the MLE of the mean parameter of the normal distribution model is equivalent to minimizing the squared error.

Why can we justify the maximum likelihood estimator (MLE)?

Similar to the sample mean, if data points are generated by a distribution indicated by a parameter vector in the parameter set of a parametric vector, the MLE has the following properties:

- ***Consistency***: The MLE converges to the true parameter.
- ***Asymptotic normality***: An appropriately scaled MLE's distribution converges to a normal distribution, and its error is proportional to $\frac{1}{\sqrt{m}}$.

Statistical test

In real applications (especially in medical applications), we need to judge from data whether a phenomenon happens or not.

Specifically, for some summary statistics θ, we want to judge from data points whether $\theta \in \mathcal{H}_1$ or not.

Statistical tests give us a framework to make such a judgement.

We cannot directly prove that “the hypothesis is correct.”

What we want to “prove” is the following statement: “if the data points’ values are $x_1, x_2, \ldots, x_m$, then $\theta \in \mathscr{H}_1$,” in some probability theory sense.

However, in (frequentism) statistics, we cannot discuss the probability of a parameter being true given data points since a parameter is not a random variable.

In contrast, we can discuss the other direction, that is, given a parameter, we can discuss the probability of the random variables taking the given values.

So, we take the **contraposition** of the proposition that we originally wanted to prove.

Null hypothesis

The contraposition of “if the data points’ values are $x_1, x_2, \ldots, x_m$, then $\theta \in \mathcal{H}_1$,” is:

“If $\theta \notin \mathcal{H}_1$, then the data points’ values are NOT $x_1, x_2, \ldots, x_m$.”

The hypothesis $\mathcal{H}_0 := \mathcal{H} \setminus \mathcal{H}_1$ is called the null hypothesis. Here, $\mathcal{H}$ is the set of all the distributions that we assume is possible as a true distribution.

Terminology: rejecting and accepting a hypothesis

- We say that we ***reject*** a hypothesis when we conclude that the true distribution is not in the hypothesis.
- We say that we ***accept*** a hypothesis when we conclude that the true distribution is not in the hypothesis.

Test statistics

We consider a summary statistic of the empirical distribution. The summary statistic is a random variable since it is a function of the data points, which are random variables. Hence, the distribution of summary statistics is determined we assume a distribution of the data points.

For a distribution, if the value of the summary statistic is unlikely taken on the distribution, then we would conclude that the data points are not generated by the distribution.

The summary statistic on which we judge which distribution the data points come from is called a ***test statistic***.

Example: are our products better?

We are going to compose a component purer than natural one. Suppose that the purity of a natural one is 92% on average.

Our factory composed a component 8 times and the purity of was the following:

Trial	1	2	3	4	5	6	7	8
Purity	95	93	94	94	92	93	91	96

8 trial results of our factory

Are our factory's products better than natural ones on average?

The sample mean of the factory's products is 93.5, which is better than 92, the natural components average. Could we conclude that our factory's products are better than natural components?

What's our concern?

The sample mean of the factory's products is 93.5, which is better than 92, the natural components average.

A possible bad story is that the true mean μ_0 is not larger than 92, but the sample mean was, luckily, 93.5, better than 92. This is our concern.

Hence, we consider how likely this bad story can happen by "luck."

t-test

Suppose that $X_1, X_2, \ldots, X_m$ are random variables independently and identically following the normal distribution with a unknown true mean μ and variance σ^2.

We want to see whether or not the true mean equals to a value μ_0. That is, the null hypothesis is $\mathscr{H}_0 = \{\mu_0\}$.

Following the idea of the statistical test, we evaluate whether or not those random variables' values are extreme under the null hypothesis $\mu = \mu_0$. For this purpose, we consider the following value, called ***t-statistic***.

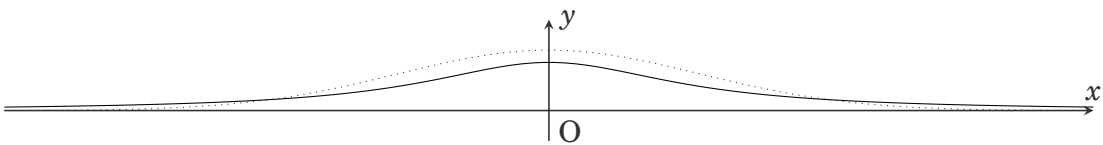
$$t := \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{m}}}, \tag{53}$$

where $\overline{X}$ and s are the sample mean and sample standard deviation defined by

$$\overline{X} := \frac{1}{m} \sum_{i=1}^{m} X_i, \quad s := \sqrt{\frac{1}{m} \left(X_i - \overline{X} \right)^2}. \tag{54}$$

t-distribution

Suppose that $X_1, X_2, \ldots, X_m$ are independently and identically following a normal distribution. Then, t follows the t-distribution with $m-1$ degree of freedom, whose PDF p_{m-1} is illustrated as follows.



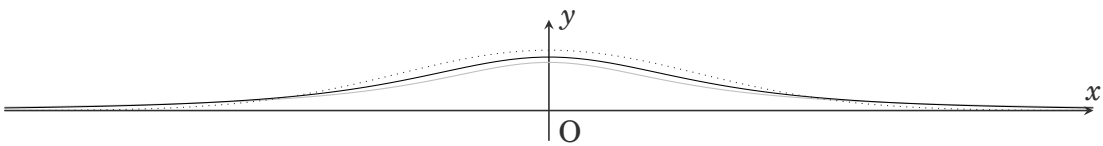
t-distributions' PDFs.

The t-distribution's PDF is symmetric and similar to the standard normal distribution's PDF but has heavier tails than a normal distribution.

As m increases, the PDF converges to the standard normal distribution's PDF.

t-distribution

Suppose that $X_1, X_2, \ldots, X_m$ are independently and identically following a normal distribution. Then, t follows the t-distribution with $m-1$ degree of freedom, whose PDF p_{m-1} is illustrated as follows.



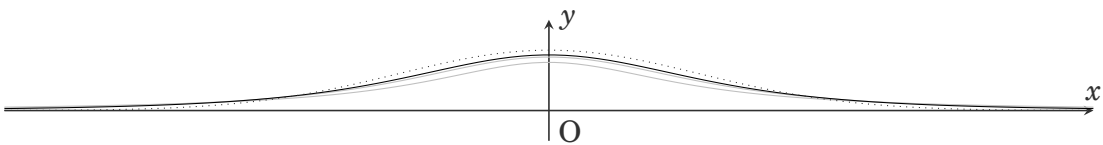
t-distributions' PDFs.

The t-distribution's PDF is symmetric and similar to the standard normal distribution's PDF but has heavier tails than a normal distribution.

As m increases, the PDF converges to the standard normal distribution's PDF.

t-distribution

Suppose that $X_1, X_2, \ldots, X_m$ are independently and identically following a normal distribution. Then, t follows the t-distribution with $m-1$ degree of freedom, whose PDF p_{m-1} is illustrated as follows.



t-distributions' PDFs.

The t-distribution's PDF is symmetric and similar to the standard normal distribution's PDF but has heavier tails than a normal distribution.

As m increases, the PDF converges to the standard normal distribution's PDF.

Note: The specific form of the t-distribution.

The PDF $p_{m-1}(x)$ of the t-distribution with $m-1$ degree of freedom is given by

$$p_{m-1}(x) = \frac{\Gamma\left(\frac{m}{2}\right)}{\sqrt{(m-1)\pi}\,\Gamma\left(\frac{m-1}{2}\right)}\left(1+\frac{x^2}{m-1}\right)^{-\frac{m}{2}} \tag{55}$$

where $\Gamma(z) := \int_0^\infty s^{z-1}\exp(-s)\mathrm{d}s$.

p-value

How do we determine the unlikeliness of the value of the test statistic?

As a criterion of the unlikeliness of the statistic's value, we consider the probability of the statistic taking a more extreme value [8]. The probability is called the ***p-value***. A small p-value indicates that the value of the statistic takes an extreme value.

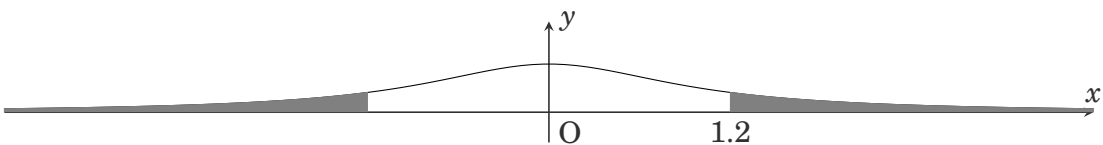
[8] Hence, we need to define in which case the value of the statistic is extreme. Although it is intuitive for well-known cases, there does not seem to be a way to mathematically decide it.

p-value in t-test

The t-statistic takes zero if $\overline{X} = \mu$. In non-extreme cases, where the sample mean $\overline{X}$ is around the mean μ, t is around zero. If extreme cases, where the sample mean $\overline{X}$ is distant from the mean μ, $|t|$ takes a large value. The larger $|t|$, the more extreme.

Here, when t-statistic takes a value t_0, we define its p-value by

$$p = \Pr(|t| > |t_0|). \tag{56}$$



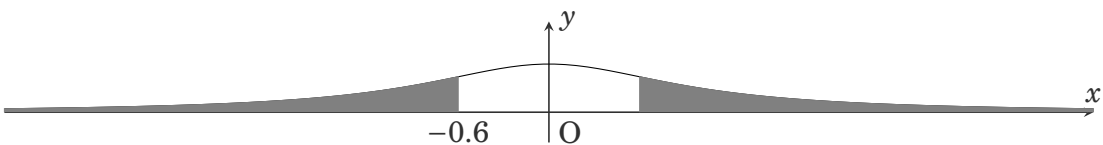
The p-value (the gray area) when t takes $t_0 = 1.2$.

p-value in t-test

The t-statistic takes zero if $\overline{X} = \mu$. In non-extreme cases, where the sample mean $\overline{X}$ is around the mean μ, t is around zero. If extreme cases, where the sample mean $\overline{X}$ is distant from the mean μ, $|t|$ takes a large value. The larger $|t|$, the more extreme.

Here, when t-statistic takes a value t_0, we define its p-value by

$$p = \Pr(|t| > |t_0|). \tag{56}$$



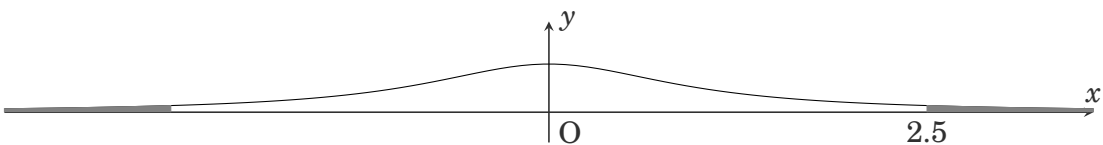
The p-value (the gray area) when t takes $t_0 = 1.2$.

p-value in t-test

The t-statistic takes zero if $\overline{X} = \mu$. In non-extreme cases, where the sample mean $\overline{X}$ is around the mean μ, t is around zero. If extreme cases, where the sample mean $\overline{X}$ is distant from the mean μ, $|t|$ takes a large value. The larger $|t|$, the more extreme.

Here, when t-statistic takes a value t_0, we define its p-value by

$$p = \Pr(|t| > |t_0|). \tag{56}$$



The p-value (the gray area) when t takes $t_0 = 1.2$.

Significance level

We reject a hypothesis consisting of a single distribution if the p-value of the distribution on the data points is small[9].

Now, how small should the threshold, called the ***significance level*** be?

There is no mathematical reason to determine it.

There is a convention to set the threshold at 0.05.

That is,

- If p-value is larger than 0.05, then we do not reject the null hypothesis $\mathscr{H}_0$.
- If p-value is smaller than 0.05, then we reject the null hypothesis and accept the alternative hypothesis $\mathscr{H}_1$.

[9] We reject a hypothesis consisting of multiple distributions if we can reject the hypothesis consisting of any distribution in the original hypothesis

t-test example

Suppose that the purity of a natural one is 92% on average.

Our factory composed a component 8 times and the purity of was the following:

Trial	1	2	3	4	5	6	7	8
Purity	95	93	94	94	92	93	91	96

8 trial results of our factory

Are our factory's products better than natural ones on average? The null hypothesis is $\mu = \mu_0 = 92$.

The sample mean and standard deviation are $\overline{X} = 93.5$ and $s \approx 1.60$. The t-statistic is $t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{m}}} \approx \frac{93.5 - 92}{\frac{1.6}{2\sqrt{2}}} = 2.65$. Here, under the null hypothesis, t follows the t-distribution with 7 degree of freedom. The p-value is $p \approx 0.032$, according to an online calculator. Since $p < 0.05$, we reject the null hypothesis, and accept the alternative hypothesis. Hence, we can statistically conclude that our factory produces better components than

False positive (Type I error) and false negative (Type II error)

Statistical tests behaves stochastically, so it may make a mistake. We may make two types of mistakes:

- ***False positive (Type I error)***: Accepts the alternative hypothesis $\mathscr{H}_1$ when the null hypothesis $\mathscr{H}_0$ is actually correct.
- ***False negative (Type II error)***: Fails to reject the null hypothesis $\mathscr{H}_0$ when the alternative hypothesis $\mathscr{H}_1$ is actually correct.

In the simple t-test case, the type I error probability equals to the significance level α.

Significance level, false-positive, false-negative

The false-positive rate, the possibility of accepting the alternative hypothesis when the data points are generated by a distribution in the null hypothesis, is determined by the significance level.

So, is it better to use a smaller significance level?

The answer is NO. It is because it increases the false-negative rate, the possibility of failing to accept the alternative hypothesis when the data points are generated by a distribution in the alternative hypothesis.