Probability Theory

SUZUKI, Atsushi

The whole contents

1. Random variables

2. Multiple Random Variables

3. Continuous Random Variables

4. Sample Statistics

5. Statistical Test

Outline

1. Random variables

    1.1 Introduction: why do we learn random variables?

    1.2 Univariate discrete random variable

    1.3 Visualization of a distribution

    1.4 Summary statistics for a univariate random variable

    1.5 Expectation

    1.6 Median

    1.7 Variance and a function of a random variable

    1.8 Exercises

Outline

Probability theory handles random events

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 variable

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

Example (RVs in real life)

- A stock price in finance
- The remainder of one's life in medicine
- The intensity of the acoustic signal in speech recognition

Random variable

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

Example (RVs in real life)

- A stock price in finance
- The remainder of one's life in medicine
- The intensity of the acoustic signal in speech recognition

One reason why we mainly discuss a random variable 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.

Learning outcomes

By the end of this topic, you should be able to:

- Explain the difference between random events and random variables,
- Represent the probability distribution of a random variable using the probability mass function and cumulative distribution function, and
- Describe a probability distribution using summary statistics.

Outline

Discrete random variable: motivation

In general, a random variable may take all the real values.

Still, when considering applications in computer science, including artificial intelligence, we do not need to handle all the real values. Specifically, we can assume that a random variable always takes a value in a finite subset of $\mathbb{R}$ (the set of real numbers).

[1] Nevertheless, we need to learn more general cases later even if we are interested in finite value cases only.

Discrete random variable: motivation

In general, a random variable may take all the real values.

Still, when considering applications in computer science, including artificial intelligence, we do not need to handle all the real values. Specifically, we can assume that a random variable always takes a value in a finite subset of $\mathbb{R}$ (the set of real numbers).

This is because a computer can handle a finite number of real numbers. For example, a computer usually uses 64 bits to represent a real value. In this case, the computer can represent only $2^{64} \approx 1.84 \times 10^{19}$ real numbers.

Hence, it is good to begin with such finite cases[1].

[1] Nevertheless, we need to learn more general cases later even if we are interested in finite value cases only.

Discrete random variables

Definition

A random variable taking a value randomly in a discrete subset [2] 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.

[2] 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.

Discrete random variables

Definition

A random variable taking a value randomly in a discrete subset [2] 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

[2] 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

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

The relation between the value that a RV takes and its probability is called the ***distribution*** of the RV. The PMF is the most fundamental way to represent the distribution of a discrete RV.

Properties of a PMF

A PMF must satisfy the following:

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

PMF tells us all we want to know.

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

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

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

A frequency is a discrete random variable

The probability theory can handle data points by considering its ***frequency***. This is the first step of ***data science***.

Suppose that we have m data points taking values in $\mathbb{R}$. For the probability theory to handle the data points, we need to construct a random variable.

Specifically, we sample a data point uniform-randomly. Then, the value of the sampled 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 of frequency

Example (Exam results)

Suppose that we have $m = 20$ students and consider their results in an exam. For $x \in \mathscr{X} = \{0, 1, 2, 3, 4, 5\}$, 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
$P_X(x) \coloneqq \Pr(X = x) = \frac{m_x}{m}$	0.15	0.10	0.15	0.25	0.30	0.05

Exam result data points and the frequency.

Outline

How to visualize a distribution?

If a distribution is complicated, then you might want to understand it from a figure, not from a long table.

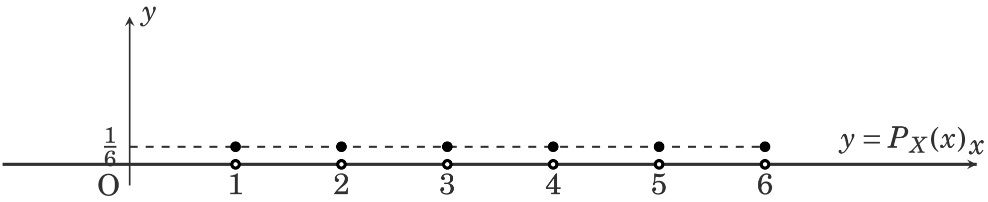
One way is to draw a graph of the PMF.

Example of a PMF graph: 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

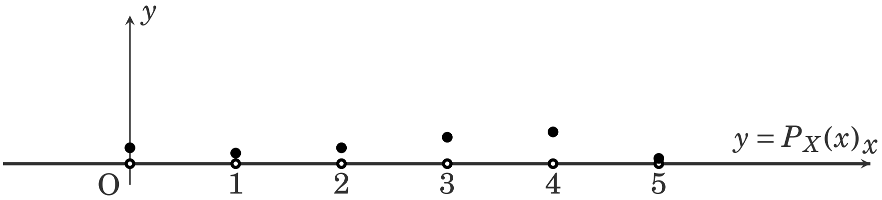


Example of a PMF graph: rolling an ideal dice

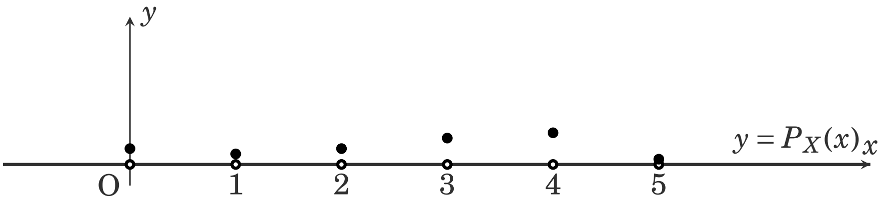
Suppose that we roll an ideal six-sided dice. The PMF is given as follows.

Score x	0	1	2	3	4	5
$P_X(x) := \Pr(X = x)$	0.15	0.10	0.15	0.25	0.30	0.05

The PMF of the frequency of exam results



Pros and cons of the PMF graph



Pros: From the PMF graph, we can easily see which value the RV takes more and less frequently.

Cons: A PMF is not suitable to calculate the probability of a RV taking a value in a certain range, e.g., $\Pr(1.5 \leq X \leq 3.8)$.

Cumulative distribution function (CDF)

Any random variable has 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

$$F_X(x) := \Pr(X \leq x). \tag{3}$$

The CDF gives formulae to evaluate a section's probability

In the following, let $a, b \in \mathbb{R}$ and $a < b$.

We have that $\Pr(X < a) = \lim_{x \nearrow a} F_X(x)$, where the right hand side is the left limit of F_X at a, given by evaluating $F_X(x - \epsilon)$ while diminishing ϵ to a positive value infinitely close to zero.

Using the above fact, we can calculate the probability of a random variable taking a value in a section using the CDF as follows.

Theorem

- $\Pr(a \leq X \leq b) = \Pr(X \leq b) - \Pr(X < a) = F_X(b) - \lim_{x \nearrow a} F_X(x)$.
- $\Pr(a < X < b) = \Pr(X < b) - \Pr(X \leq a) = \lim_{x \nearrow b} F_X(x) - F_X(a)$.
- $\Pr(a < X \leq b) = \Pr(X \leq b) - \Pr(X \leq a) = F_X(b) - F_X(a)$.
- $\Pr(a \leq X < b) = \Pr(X < b) - \Pr(X < a) = \lim_{x \nearrow b} F_X(x) - \lim_{x \nearrow a} F_X(x)$.

Example of CDF: rolling an ideal dice

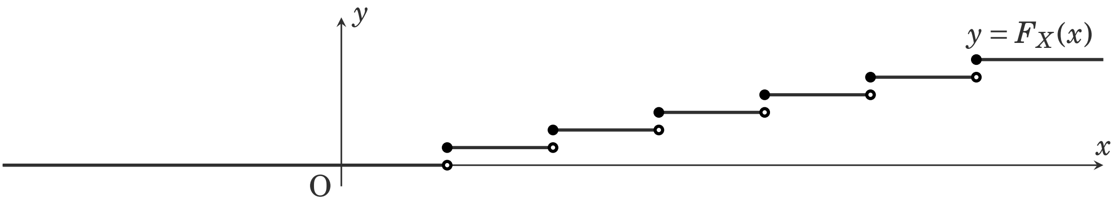
Suppose that we roll an ideal six-sided dice. The PMF is given as follows.

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

Example of CDF: rolling an ideal dice

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

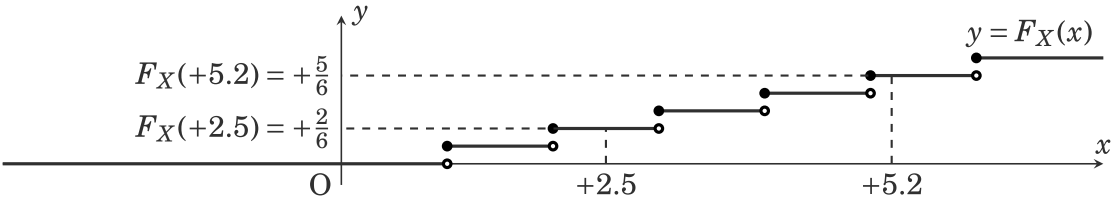


x	$(-\infty, 1)$	$[1, 2)$	$[2, 3)$	$[3, 4)$	$[4, 5)$	$[5, 6)$	$[6, +\infty)$
$F_X(x) := \Pr(X = x)$	0	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1

The CDF of rolling an ideal six-sided dice

Example of CDF: rolling an ideal dice

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

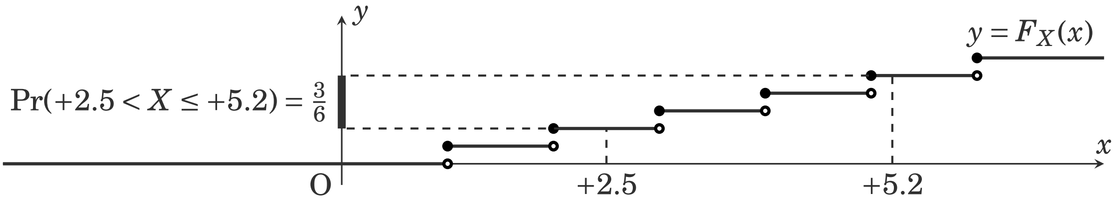


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned} \Pr(+2.5 < X \leq +5.2) &= \Pr(X \leq +5.2) - \Pr(X \leq +2.5) \\ &= F_X(+5.2) - F_X(+2.5) \\ &= \frac{5}{6} - \frac{2}{6} = \frac{3}{6}. \end{aligned} \tag{4}$$

Example of CDF: rolling an ideal dice

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

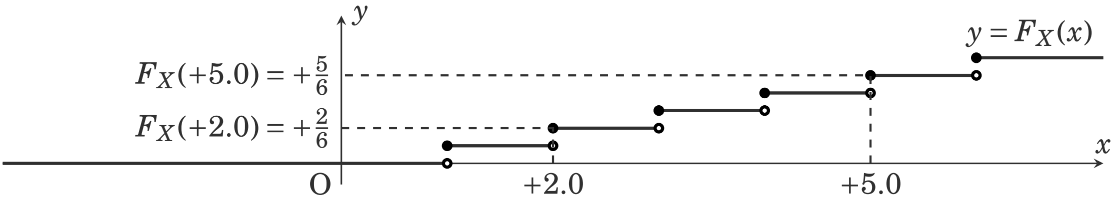


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned}\Pr(+2.5 < X \le +5.2) &= \Pr(X \le +5.2) - \Pr(X \le +2.5) \\ &= F_X(+5.2) - F_X(+2.5) \\ &= \frac{5}{6} - \frac{2}{6} = \frac{3}{6}.\end{aligned} \tag{4}$$

Example of CDF: rolling an ideal dice

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

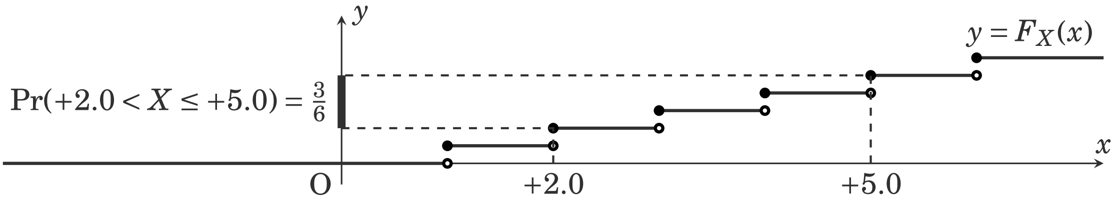


Using the CDF, we can calculate the probability of various events. For example,

$$
\begin{aligned}
\Pr(+2.0 < X \leq +5.0) &= \Pr(X \leq +5.0) - \Pr(X \leq +2.0) \\
&= F_X(+5.0) - F_X(+2.0) \\
&= \frac{5}{6} - \frac{2}{6} = \frac{3}{6}.
\end{aligned} \tag{4}
$$

Example of CDF: rolling an ideal dice

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

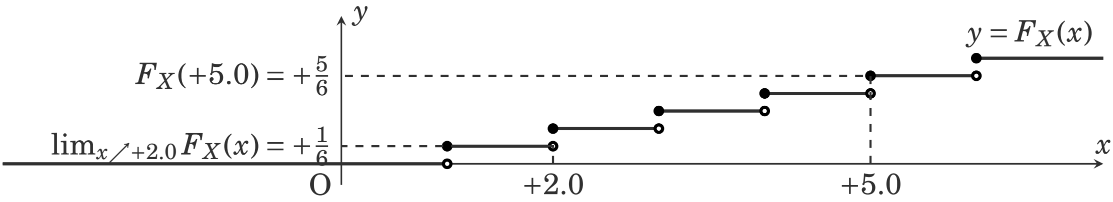


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned}\Pr(+2.0 < X \leq +5.0) &= \Pr(X \leq +5.0) - \Pr(X \leq +2.0)\\ &= F_X(+5.0) - F_X(+2.0)\\ &= \frac{5}{6} - \frac{2}{6} = \frac{3}{6}.\end{aligned} \tag{4}$$

Example of CDF: rolling an ideal dice

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

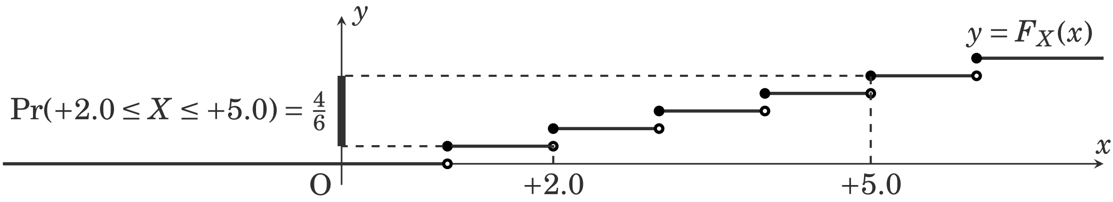


Using the CDF, we can calculate the probability of various events. For example,

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

Example of CDF: rolling an ideal dice

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

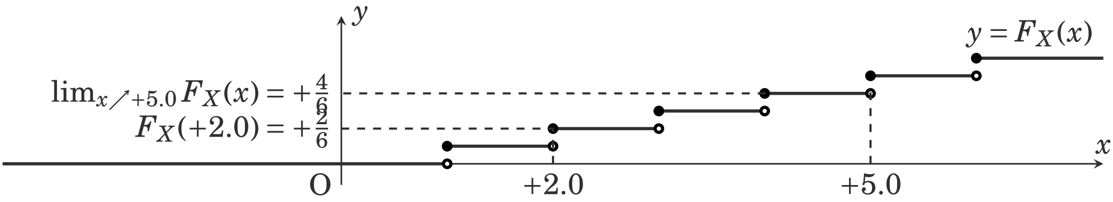


Using the CDF, we can calculate the probability of various events. For example,

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

Example of CDF: rolling an ideal dice

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

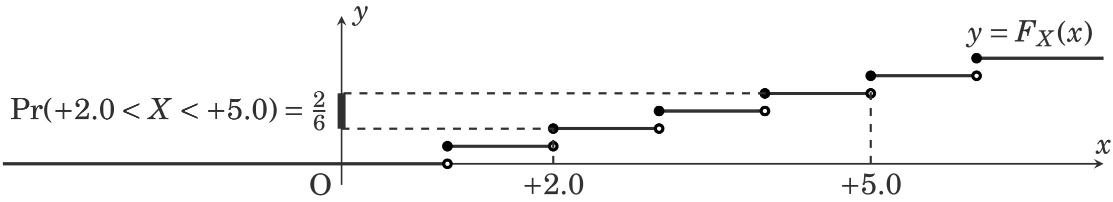


Using the CDF, we can calculate the probability of various events. For example,

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

Example of CDF: rolling an ideal dice

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

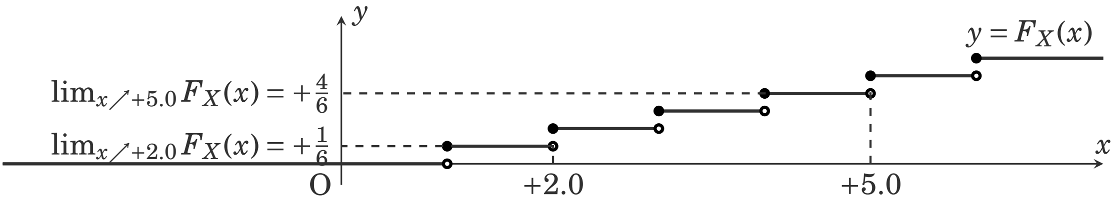


Using the CDF, we can calculate the probability of various events. For example,

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

Example of CDF: rolling an ideal dice

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

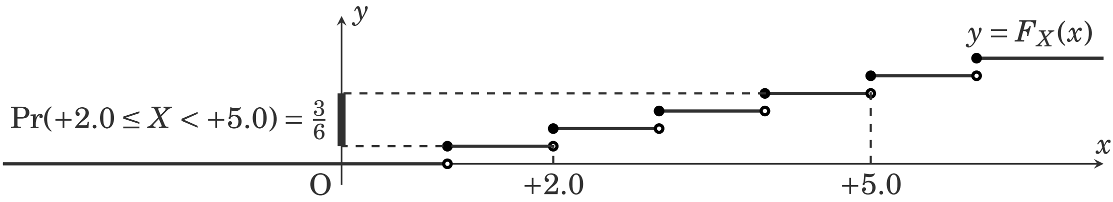


Using the CDF, we can calculate the probability of various events. For example,

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

Example of CDF: 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,

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

Example of CDF: student score frequency

Suppose that X is a random variable whose PMF is given as follows.

x	0	1	2	3	4	5
$P_X(x) := \Pr(X = x)$	0.15	0.10	0.15	0.25	0.30	0.05

The PMF of a student exam result frequency

The CDF is given as the cumulative sum of the PMF, as follows.

x	$(-\infty, 0)$	$[0, 1)$	$[1, 2)$	$[2, 3)$	$[3, 4)$	$[4, 5)$	$[5, +\infty)$
$F_X(x) := \Pr(X = x)$	0.00	0.15	0.25	0.40	0.65	0.95	1.00

The CDF of a student exam result frequency

Example of CDF: student score frequency

The CDF is given as the cumulative sum of the PMF, as follows.

x	$(-\infty, 0)$	$[0, 1)$	$[1, 2)$	$[2, 3)$	$[3, 4)$	$[4, 5)$	$[5, +\infty)$
$F_X(x) := \Pr(X = x)$	0.00	0.15	0.25	0.40	0.65	0.95	1.00

The CDF of a student exam result frequency

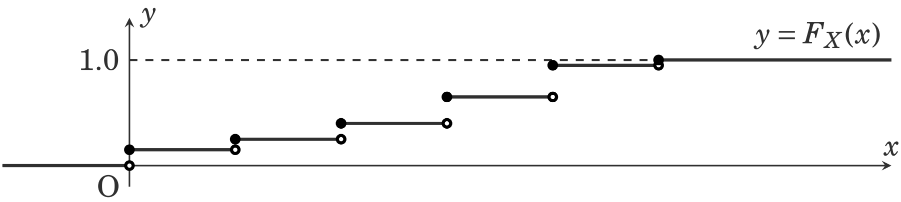
Example of CDF: student score frequency

The CDF is given as the cumulative sum of the PMF, as follows.

x	$(-\infty, 0)$	$[0, 1)$	$[1, 2)$	$[2, 3)$	$[3, 4)$	$[4, 5)$	$[5, +\infty)$
$F_X(x) := \Pr(X = x)$	0.00	0.15	0.25	0.40	0.65	0.95	1.00

The CDF of a student exam result frequency

The graph of the CDF is as follows.



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., $\lim_{x \searrow x_0} F_X(x) = F_X(x_0)$ for all $x_0 \in \mathbb{R}$.
- The CDF has its left-limit $\lim_{x \nearrow x_0} F_X(x)$ for all $x_0 \in \mathbb{R}$.

Appendix: the definition of the left limit

Definition (left/right limit/continuous)

Let $f : \mathbb{R} \to \mathbb{R}$ be a real function and a be a real value.
Suppose that for all $\delta > 0$ there exists $\epsilon > 0$ such that $\left| f\left(a - \epsilon'\right) - c \right| < \delta$ for all ϵ' that satisfies $0 < \epsilon' < \epsilon$.
Then the value c is called the ***left limit*** of a function f at $a \in \mathbb{R}$ and denoted by $\lim_{x \nearrow a} f(x)$.
We have the definition of the ***right limit*** by replacing $\left(a - \epsilon'\right)$ with $\left(a + \epsilon'\right)$ in the definition of the left limit.
The right limit is denoted by $\lim_{x \searrow a} f(x)$.
A function f is called ***left continuous*** at $a \in \mathbb{R}$ if $\lim_{x \nearrow a} f(x) = f(a)$ and ***right continuous*** at $a \in \mathbb{R}$ if $\lim_{x \searrow a} f(x) = f(a)$.
If a function is left/right continuous at every value in its domain, then we simply call the function left/right continuous.

Appendix: relation between the limit and the left and right limits

Theorem

Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a real function and a be a real value.

- $\lim_{x\to a} f(x) = c$ *if and only if* $\lim_{x\nearrow a} f(x) = \lim_{x\searrow a} f(x) = c$.
- f *is continuous at* a *if and only if* f *is left continuous and right continuous at* a.

Outline

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***[3].

[3] 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.

Outline

Definition of expectation (mean)

The most fundamental central tendency measure of a distribution is the ***expectation***.

Definition (Expectation of a discrete RV)

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

$$\mathbb{E}X := \sum_{x \in \mathscr{X}} x P_X(x). \tag{5}$$

The expectation is also called the ***mean***. Indeed, if the probability distribution is a frequency of data points, the expectation is nothing but the mean of the data points.

Example of expectation calculation

Suppose that X is a random variable whose PMF P_X is given by the following table.

x	-2	-1	0	$+1$	$+2$
$P_X(x)$	0.05	0.10	0.20	0.10	0.55

Example random function and its PMF.

We can calculate the expectation $\mathbb{E}X$ by the following procedure.

- **Step 1**:
- **Step 2**:

Example of expectation calculation

Suppose that X is a random variable whose PMF P_X is given by the following table.

x	-2	-1	0	$+1$	$+2$
$P_X(x)$	0.05	0.10	0.20	0.10	0.55
$xP_X(x)$					

Example random function and its PMF.

We can calculate the expectation $\mathbb{E}X$ by the following procedure.

- **Step 1**: Calculate $xP_X(x)$ for each $x \in \mathscr{X}$.
- **Step 2**:

Example of expectation calculation

Suppose that X is a random variable whose PMF P_X is given by the following table.

x	-2	-1	0	$+1$	$+2$
$P_X(x)$	0.05	0.10	0.20	0.10	0.55
$xP_X(x)$	-0.10				

Example random function and its PMF.

We can calculate the expectation $\mathbb{E}X$ by the following procedure.

- **Step 1**: Calculate $xP_X(x)$ for each $x \in \mathcal{X}$.
- **Step 2**:

Example of expectation calculation

Suppose that X is a random variable whose PMF P_X is given by the following table.

x	-2	-1	0	$+1$	$+2$
$P_X(x)$	0.05	0.10	0.20	0.10	0.55
$xP_X(x)$	-0.10	-0.10			

Example random function and its PMF.

We can calculate the expectation $\mathbb{E}X$ by the following procedure.

- **Step 1**: Calculate $xP_X(x)$ for each $x \in \mathscr{X}$.
- **Step 2**:

Example of expectation calculation

Suppose that X is a random variable whose PMF P_X is given by the following table.

x	-2	-1	0	$+1$	$+2$
$P_X(x)$	0.05	0.10	0.20	0.10	0.55
$xP_X(x)$	-0.10	-0.10	0.00		

Example random function and its PMF.

We can calculate the expectation $\mathbb{E}X$ by the following procedure.

- **Step 1**: Calculate $xP_X(x)$ for each $x \in \mathscr{X}$.
- **Step 2**:

Example of expectation calculation

Suppose that X is a random variable whose PMF P_X is given by the following table.

x	-2	-1	0	$+1$	$+2$
$P_X(x)$	0.05	0.10	0.20	0.10	0.55
$xP_X(x)$	-0.10	-0.10	0.00	$+0.10$	$+1.10$

Example random function and its PMF.

We can calculate the expectation $\mathbb{E}X$ by the following procedure.

- **Step 1**: Calculate $xP_X(x)$ for each $x \in \mathscr{X}$.
- **Step 2**:

Example of expectation calculation

Suppose that X is a random variable whose PMF P_X is given by the following table.

x	-2	-1	0	$+1$	$+2$
$P_X(x)$	0.05	0.10	0.20	0.10	0.55
$xP_X(x)$	-0.10	-0.10	0.00	$+0.10$	$+1.10$

Example random function and its PMF.

We can calculate the expectation $\mathbb{E}X$ by the following procedure.

- **Step 1**: Calculate $xP_X(x)$ for each $x \in \mathscr{X}$.
- **Step 2**: Evaluate the sum $\sum_{x \in \mathscr{X}} xP_X(x)$, which equals the expectation $\mathbb{E}X$.

Example of expectation calculation

Suppose that X is a random variable whose PMF P_X is given by the following table.

x	-2	-1	0	$+1$	$+2$
$P_X(x)$	0.05	0.10	0.20	0.10	0.55
$xP_X(x)$	-0.10	-0.10	0.00	$+0.10$	$+1.10$

Example random function and its PMF.

We can calculate the expectation $\mathbb{E}X$ by the following procedure.

- **Step 1**: Calculate $xP_X(x)$ for each $x \in \mathscr{X}$.
- **Step 2**: Evaluate the sum $\sum_{x \in \mathscr{X}} xP_X(x)$, which equals the expectation $\mathbb{E}X$.
  In the above case, the expectation $\mathbb{E}X$ is given by
  $\mathbb{E}X = (-0.10) + (-0.10) + 0.00 + 0.10 + 1.10 = 1.00.$

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.

The expectation is easily “warped” by outliers.

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

Example (Imbalanced score distribution)

Suppose you got a score of 99 in an exam where 100 students participated and the expectation was 98, you might feel you did very well.
However, it might be just that one student who got a score of 1 decreased the expectation significantly, as follows.

Score x	1	99	100
# students m_x	1	1	98
$P_X := \Pr(X = x) = \frac{m_x}{m}$	0.01	0.01	0.98

Exam result data points and the frequency.

Outline

Median’s idea

If a random value takes an extremely large or small value in a small probability, some might want to use the ***median*** as a summary statistic.

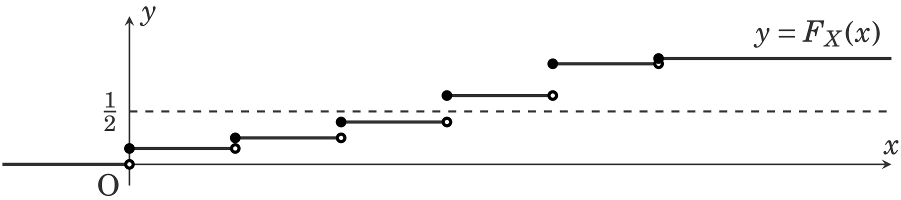
Roughly speaking, the median is defined so that the random variable is larger than the median in 50% probability and smaller than the median in 50% probability.

Median's idea

If a random value takes an extremely large or small value in a small probability, 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.

In other words, the median is the value x such that the graph $y = F_X(x)$ of the CDF crosses the horizontal line $y = \frac{1}{2}$.

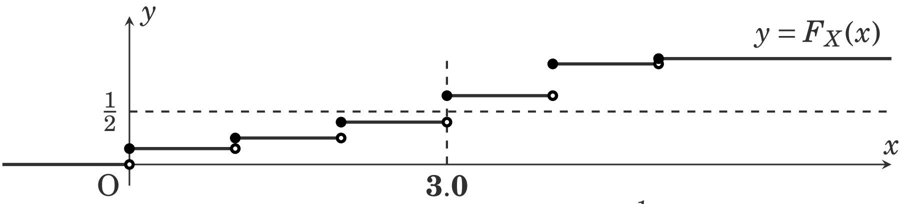


Median's idea

If a random value takes an extremely large or small value in a small probability, 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.

In other words, the median is the value x such that the graph $y = F_X(x)$ of the CDF crosses the horizontal line $y = \frac{1}{2}$.



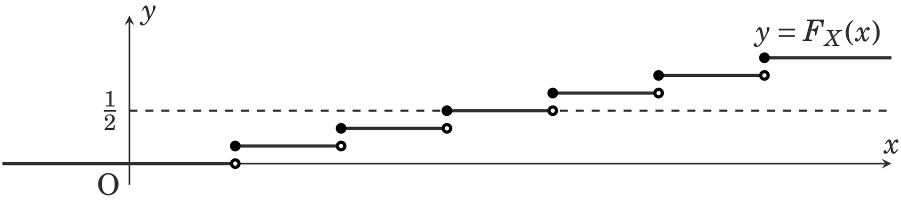
where the $y = F(x)$ crosses $y = \frac{1}{2}$

Median’s idea

If a random value takes an extremely large or small value in a small probability, 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.

If the CDF graph has a horizontal segment on $y = \frac{1}{2}$, the median is the middle point of the segment.

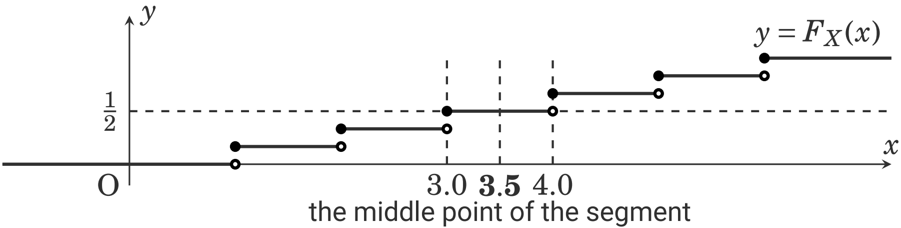


Median's idea

If a random value takes an extremely large or small value in a small probability, 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.

If the CDF graph has a horizontal segment on $y = \frac{1}{2}$, the median is the middle point of the segment.



Definition of median

Definition (The 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:

$$\Pr(X \le M) \ge \frac{1}{2} \text{ and } \Pr(X \ge M) \ge \frac{1}{2}. \tag{6}$$

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

The definition of the median

Definition (The definition of the median)

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

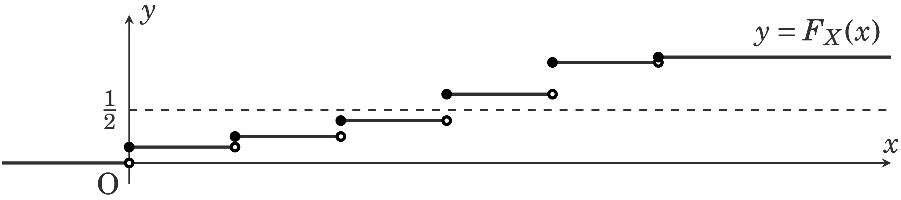
$$\begin{aligned} \underline{M} &:= \min\left\{ M \in \mathbb{R} \middle| \Pr(X \leq M) \geq \frac{1}{2} \text{ and } \Pr(X \geq M) \geq \frac{1}{2} \right\}, \\ \overline{M} &:= \max\left\{ M \in \mathbb{R} \middle| \Pr(X \leq M) \geq \frac{1}{2} \text{ and } \Pr(X \geq M) \geq \frac{1}{2} \right\}. \end{aligned} \tag{7}$$

The ***median*** M is defined as the midpoint of $\underline{M}$ and $\overline{M}$, i.e., $M := \frac{\underline{M}+\overline{M}}{2}$.

The above definition looks complicated, but it is in fact easy if we see the CDF graph.

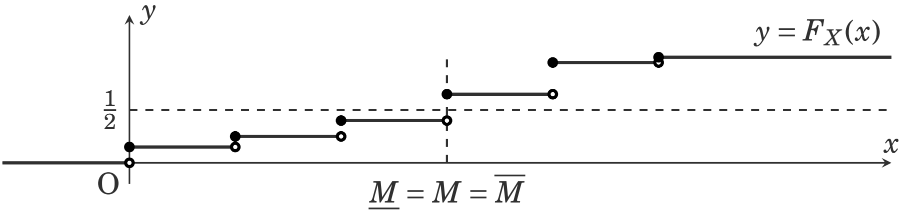
The definition of the median by the CDF graph

If the CDF graph “crosses” the graph of $y = \frac{1}{2}$,



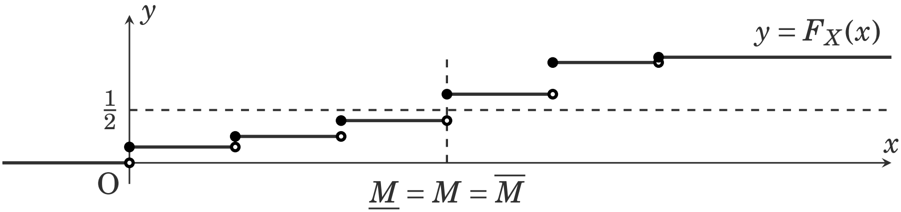
The definition of the median by the CDF graph

If the CDF graph "crosses" the graph of $y = \frac{1}{2}$,

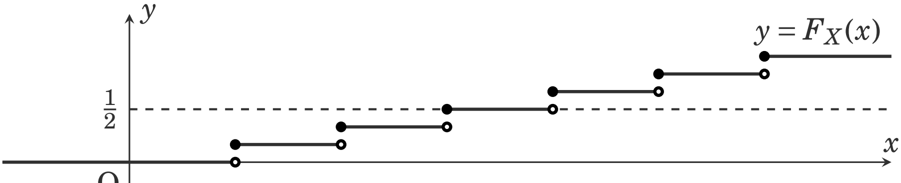


The definition of the median by the CDF graph

If the CDF graph "crosses" the graph of $y = \frac{1}{2}$,

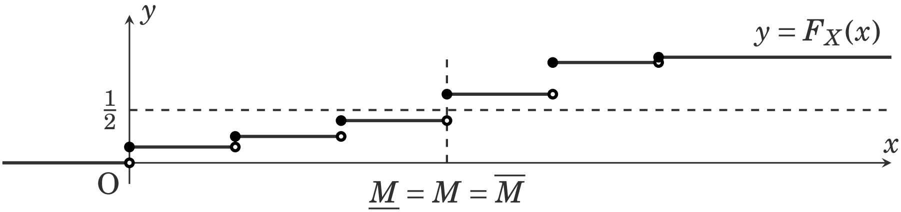


If the CDF graph has a horizontal segment on $y = \frac{1}{2}$,

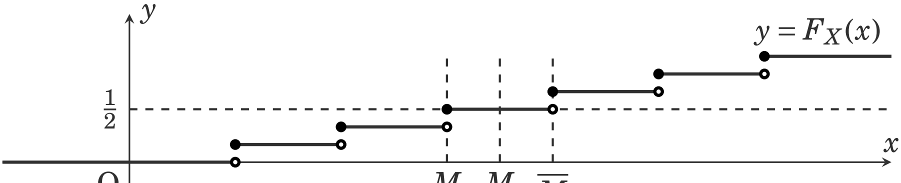


The definition of the median by the CDF graph

If the CDF graph "crosses" the graph of $y = \frac{1}{2}$,



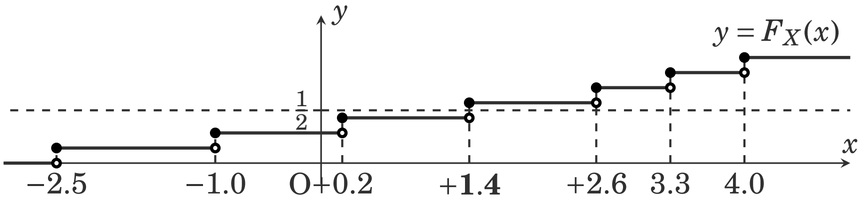
If the CDF graph has a horizontal segment on $y = \frac{1}{2}$,



Median of frequency for an odd data point case

By definition, the median of the frequency of $(2k+1)$ data points is the value of the $(k+1)$th largest data point. This is equivalent to the $(k+1)$th smallest data point. In this sense, the definition is symmetric. The value is simply called the median of the data points.

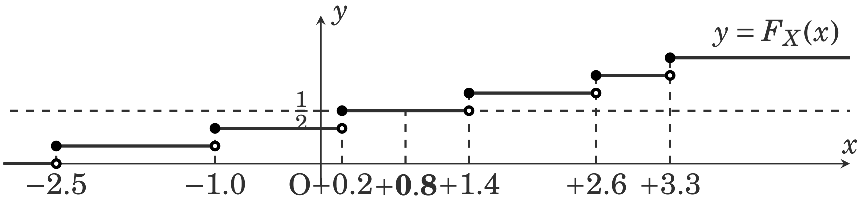
For example, if we have 7 sorted data points $(-2.5, -1.0, +0.2, +1.4, +2.6, +3.3, +4.0)$, then the median is the value of the 4th largest (or equivalently, the 4th smallest) data point, which is $+1.4$.



Median of frequency for an even data point case

By definition, the median of the frequency of $2k$ datapoints is the middle point of the values of the kth and $k+1$th largest data points. This is equivalent to the middle point of the values of the kth and $k+1$th largest data points. In this sense, the definition is symmetric. The value is simply called the median of the data points.

For example, if we have 6 sorted data points $(-2.5, -1.0, 0.2, +1.4, +2.6, +3.5)$, then the median is the middle point of the values of the 3rd and 4th largest (or equivalently, the 3rd and 4th smallest) data points, which is $\frac{0.2+1.4}{2} = 0.8$.



Median of imbalanced data

Example

Consider the following exam results of 100 participants given by the following table and the frequency of the data points.

Score x	1	99	100
# students m_x	1	1	98
$P_X := \Pr(X = x) = \frac{m_x}{m}$	0.01	0.01	0.98

Exam result data points and the frequency.

Since we have 100 students, which is an even number, the median is the middle point of the 50th-best student's score and the 51th-best student's score, which is 100.

Median tends to ignore “minor” data points

It is not that the median is a perfect statistic. Indeed, the median tends to ignore a relatively minor cohort even though the size of the cohort is not ignorable.

Example

Consider the following exam results of 100 participants given by the following table and the frequency of the data points.

Score x	0	100
# students m_x	49	51
$P_X := \Pr(X = x) = \frac{m_x}{m}$	0.49	0.51

Exam result data points and the frequency.

Then, the median is the middle point of the 50th-best student’s score and the 51th-best student’s score, which is 100. However, this median ignores the 49%, who received zero scores.

Outline

The basic idea of variance as a variability measure

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 expect to create a variability measure using $\left(X-\mu\right)^2$.

But, what is $\left(X-\mu\right)^2$?

The basic idea of variance as a variability measure

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 $(X-\mu)^2$. If X is far (whether large or not) from μ, the square deviation $(X-\mu)^2$ is large.

Hence, we expect to create a variability measure using $(X-\mu)^2$.

But, what is $(X-\mu)^2$? Since it depends on the value of X, $(X-\mu)^2$ is (the output value of) a function of X, and since X is a random variable, $(X-\mu)^2$ is **also a random variable**!

The ***variance*** is nothing but the expectation of the RV $(X-\mu)^2$. To understand this amount, let’s discuss the function of random variables in general.

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.

In particular, if X is a discrete RV, then $f(X)$ is also a discrete RV. Specifically, if the support of X is $\mathscr{X}$, then the support of $f(X)$ is $\{f(x) | x \in \mathscr{X}\}$.

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

Example of a function of a RV

Define f by $f(x) = x^2$, and suppose 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.

Suppose that we are interested in the behavior of $f(X)$. The variable $f(X)$ is also a random variable since it depends on the random behavior of the RV X.

Now, what are the support, the PMF, and the expectation of $f(X) = X^2$?

Example of a function of a RV

Define f by $f(x) = x^2$, and suppose 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.

Let's find the **support** of $f(X) = X^2$. The RV X takes a value in $\mathscr{X} = \{-1, 0, +1\}$. Since $f(-1) = (-1)^2 = +1$, $f(0) = (0)^2 = 0$, and $f(+1) = (+1)^2 = +1$, The RV $f(X) = X^2$ only takes a value 0 or $+1$ only. Hence the support of $f(X) = X^2$ is $\{0, +1\}$. In particular, $f(X)$ is also a discrete RV.

Example of a function of a RV

Define f by $f(x) = x^2$, and suppose 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.

Let's find the **PMF** P_{X^2}.

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

Since $X^2 = 0 \Leftrightarrow X = 0$ holds,[4] 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 $X^2 = 0$.

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

Example of a function of a RV

Define f by $f(x) = x^2$, and suppose 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.

Let's find the **PMF** 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

$$\begin{aligned}\Pr\left(X^2 = 1\right) &= \Pr(\text{“}X = -1 \text{ or } X = -1\text{”}) \\ &= \Pr(X = -1) + \Pr(X = +1) = 0.2 + 0.5 = 0.7.\end{aligned} \tag{8}$$

Here, the second equation comes from the sum law since "$X = -1$ and $X = -1$" do not happen at the same time.

Example of a function of a RV

Define f by $f(x) = x^2$, and suppose 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.

To wrap up,

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

The PMF of X^2.

Example of a function of a RV

Define f by $f(x) = x^2$, and suppose 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.

Let's evaluate the **expectation** $\mathbb{E}f(X) = \mathbb{E}X^2$.

Example of a function of a RV

Define f by $f(x) = x^2$, and suppose 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.

Let's evaluate the **expectation** $\mathbb{E}f(X) = \mathbb{E}X^2$. If we use the PMF of X^2, it looks like

$$\begin{aligned} \mathbb{E}X^2 &= 0 \cdot P_{X^2}(0) + (+1) \cdot P_{X^2}(+1) \\ &= 0 \cdot 0.3 + (+1) \cdot 0.7. \end{aligned} \tag{8}$$

Example of a function of a RV

Define f by $f(x) = x^2$, and suppose 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.

Let's evaluate the **expectation** $\mathbb{E} f(X) = \mathbb{E} X^2$. Since $P_{X^2}(0)$ equals $P_X(0)$ and $P_{X^2}(+1)$ equals the sum $P_{X^2}(-1) + P_{X^2}(+1)$, we can rewrite it using P_X only.

$$\begin{aligned} \mathbb{E} X^2 &= 0 \cdot P_{X^2}(0) + (+1) \cdot P_{X^2}(+1) \\ &= 0^2 \cdot P_X(0) + \left[(-1)^2 \cdot P_X(-1) + (+1)^2 \cdot P_X(+1)\right] \\ &= \sum_{x \in \{-1,0,+1\}} f(x) P_X(x) \end{aligned} \tag{8}$$

Behaviors of A function of a RV

If we generalize the previous discussion, we have the following theorem.

Theorem

Suppose that X is a RV and $f : \mathbb{R} \to \mathbb{R}$ are a real-valued function taking a real variable as an input. Then, $f(X)$ is also a RV.
In particular, if X is a discrete RV, $f(X)$ is also a discrete RV. Furthermore, if the support and PMF of X are denoted by $\mathscr{X}$ and P_X, respectively,

- *The support of $f(X)$ is $\{f(x) | x \in \mathscr{X}\}$,*
- *The PMF $P_{f(X)}$ is given by $P_{f(X)}(y) = \sum_{x \in \{x' | f(x') = y\}} P_X(x)$,*
- *The expectation $\mathbb{E} f(X)$ is given by $\mathbb{E} f(X) = \sum_{x \in \mathscr{X}} f(x) P_X(x)$.*

The linearity of the expectation

The expectation operator $\mathbb{E}$ has the property called ***linearity***, which often makes the expectation calculation of a complicated function easier.

Theorem (The linearity of the expectation)

Let X be a random variable, $a, b \in \mathbb{R}$ be real numbers, and $f, g : \mathbb{R} \to \mathbb{R}$ be real-valued functions taking a real variable. Then, we have that

$$\mathbb{E}[af(X) + bg(X)] = a\mathbb{E}f(X) + b\mathbb{E}g(X). \tag{9}$$

The above theorem provides us with the formula for the expectation calculation of a linear function of a RV.

Corollary

Let X be a random variable and $a, b \in \mathbb{R}$ be real numbers. Then, we have that

$$\mathbb{E}[aX + b] = a\mathbb{E}X + b. \tag{10}$$

Example of a linear function's expectation

Example

Suppose that X is a random variable whose PMF P_X is given by the following table.

x	-2	-1	0	$+1$	$+2$
$P_X(x)$	0.05	0.10	0.20	0.10	0.55

Example random function and its PMF.

The expectation is given by $\mathbb{E}X = 1.00$.
Let's consider the random function given by $-3X+5$ and its expectation.
According to the formula, $\mathbb{E}[-3X+5] = -3\,\mathbb{E}X + 5 = (-3)\cdot 1.00 + 5 = 2.00$.
Note that the PMF P_{-3X+5} is given by the following, which we did not use to calculate $\mathbb{E}[-3X+5]$.

x	$+11$	$+8$	$+5$	$+2$	-1
$P_{-3X+5}(x)$	0.05	0.10	0.20	0.10	0.55

The PMF of $-3X+5$.

Proof: the linearity of the expectation

Proof.

$$\begin{aligned} \mathbb{E}[af(X)+bg(X)] &= \sum_{x\in\mathcal{X}} [af(x)+bg(x)]P_X(x) \\ &= a\sum_{x\in\mathcal{X}} f(x)P_X(x) + b\sum_{x\in\mathcal{X}} g(x)P_X(x) \\ &= a\,\mathbb{E}f(X) + b\,\mathbb{E}g(X). \end{aligned} \tag{11}$$

$\square$

Definition of variance

Recall the basic idea of the variance.

Let X be a random variable and μ be its expectation. The ***square deviation*** of X is defined as $(X-\mu)^2$. If X is far (whether large or not) from μ, the square deviation $(X-\mu)^2$ 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 [4] $(X-\mu)^2$, that is,

$$\mathbb{V}[X] := \mathbb{E}(X-\mu)^2. \tag{12}$$

[4] 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 variance

Recall that the variance is defined by $\mathbb{V}[X] := \mathbb{E}\left(X - \mu\right)^2$. Using the formula to calculate the expectation of the discrete RV, we get the following formula to calculate the variance of a discrete random variable.

Theorem

Let X be a discrete random variable taking values in $\mathscr{X} \subset \mathbb{R}$. Also, suppose that $\mu := \mathbb{E}X$ and $P_X : \mathscr{X} \to [0,1]$ are its expectation and PMF, respectively.
The variance $\mathbb{V}[X]$ is given by

$$\mathbb{V}[X] = \sum_{x \in \mathscr{X}} \left(x - \mu\right)^2 P_X(x). \tag{13}$$

Example of variance calculation

x	-2	-1	0	$+1$	$+2$
Probability mass $P_X(x)$	0.05	0.10	0.20	0.10	0.55
Deviation $x - \mu_x$					
Square deviation $\left(x - \mu_X\right)^2$					
Weighted sq. dev. $\left(x - \mu_X\right)^2 P_X(x)$					

Example random function and its PMF.

- **Step 1**: Calculate the expectation $\mu_X = \mathbb{E}X$ of X.
  In the above example, we have $\mu_X = \mathbb{E}X = +1.00$.
- **Step 2**:
- **Step 3**:

Example of variance calculation

x	-2	-1	0	$+1$	$+2$
Probability mass $P_X(x)$	0.05	0.10	0.20	0.10	0.55
Deviation $x - \mu_x$					
Square deviation $\left(x - \mu_X\right)^2$					
Weighted sq. dev. $\left(x - \mu_X\right)^2 P_X(x)$					

Example random function and its PMF.

- **Step 1**: Calculate the expectation $\mu_X = \mathbb{E}X$ of X.
  In the above example, we have $\mu_X = \mathbb{E}X = +1.00$.
- **Step 2**: Calculate the deviation $x - \mu_X$, the square deviation $\left(x - \mu_X\right)^2$, and the weighted square deviation $\left(x - \mu_X\right)^2 P_X(x)$ for every $x \in \mathscr{X}$.
- **Step 3**:

Example of variance calculation

x	-2	-1	0	$+1$	$+2$
Probability mass $P_X(x)$	0.05	0.10	0.20	0.10	0.55
Deviation $x - \mu_x$	-3.00				
Square deviation $\left(x - \mu_X\right)^2$					
Weighted sq. dev. $\left(x - \mu_X\right)^2 P_X(x)$					

Example random function and its PMF.

- **Step 1**: Calculate the expectation $\mu_X = \mathbb{E}X$ of X.
  In the above example, we have $\mu_X = \mathbb{E}X = +1.00$.
- **Step 2**: Calculate the deviation $x - \mu_X$, the square deviation $\left(x - \mu_X\right)^2$, and the weighted square deviation $\left(x - \mu_X\right)^2 P_X(x)$ for every $x \in \mathscr{X}$.
- **Step 3**:

Example of variance calculation

x	-2	-1	0	$+1$	$+2$
Probability mass $P_X(x)$	0.05	0.10	0.20	0.10	0.55
Deviation $x-\mu_x$	-3.00				
Square deviation $\left(x-\mu_X\right)^2$	9.00				
Weighted sq. dev. $\left(x-\mu_X\right)^2 P_X(x)$					

Example random function and its PMF.

- **Step 1**: Calculate the expectation $\mu_X = \mathbb{E}X$ of X.
  In the above example, we have $\mu_X = \mathbb{E}X = +1.00$.
- **Step 2**: Calculate the deviation $x-\mu_X$, the square deviation $\left(x-\mu_X\right)^2$, and the weighted square deviation $\left(x-\mu_X\right)^2 P_X(x)$ for every $x \in \mathscr{X}$.
- **Step 3**:

Example of variance calculation

x	-2	-1	0	$+1$	$+2$
Probability mass $P_X(x)$	0.05	0.10	0.20	0.10	0.55
Deviation $x - \mu_x$	-3.00				
Square deviation $(x - \mu_X)^2$	9.00				
Weighted sq. dev. $(x - \mu_X)^2 P_X(x)$	0.45				

Example random function and its PMF.

- **Step 1**: Calculate the expectation $\mu_X = \mathbb{E}X$ of X.
  In the above example, we have $\mu_X = \mathbb{E}X = +1.00$.
- **Step 2**: Calculate the deviation $x - \mu_X$, the square deviation $(x - \mu_X)^2$, and the weighted square deviation $(x - \mu_X)^2 P_X(x)$ for every $x \in \mathscr{X}$.
- **Step 3**:

Example of variance calculation

x	-2	-1	0	$+1$	$+2$
Probability mass $P_X(x)$	0.05	0.10	0.20	0.10	0.55
Deviation $x - \mu_x$	-3.00	-2.00			
Square deviation $\left(x - \mu_X\right)^2$	9.00				
Weighted sq. dev. $\left(x - \mu_X\right)^2 P_X(x)$	0.45				

Example random function and its PMF.

- **Step 1**: Calculate the expectation $\mu_X = \mathbb{E}X$ of X.
  In the above example, we have $\mu_X = \mathbb{E}X = +1.00$.
- **Step 2**: Calculate the deviation $x - \mu_X$, the square deviation $\left(x - \mu_X\right)^2$, and the weighted square deviation $\left(x - \mu_X\right)^2 P_X(x)$ for every $x \in \mathscr{X}$.
- **Step 3**:

Example of variance calculation

x	-2	-1	0	$+1$	$+2$
Probability mass $P_X(x)$	0.05	0.10	0.20	0.10	0.55
Deviation $x - \mu_x$	-3.00	-2.00			
Square deviation $\left(x - \mu_X\right)^2$	9.00	4.00			
Weighted sq. dev. $\left(x - \mu_X\right)^2 P_X(x)$	0.45				

Example random function and its PMF.

- **Step 1**: Calculate the expectation $\mu_X = \mathbb{E}X$ of X.
  In the above example, we have $\mu_X = \mathbb{E}X = +1.00$.
- **Step 2**: Calculate the deviation $x - \mu_X$, the square deviation $\left(x - \mu_X\right)^2$, and the weighted square deviation $\left(x - \mu_X\right)^2 P_X(x)$ for every $x \in \mathscr{X}$.
- **Step 3**:

Example of variance calculation

x	-2	-1	0	$+1$	$+2$
Probability mass $P_X(x)$	0.05	0.10	0.20	0.10	0.55
Deviation $x-\mu_x$	-3.00	-2.00			
Square deviation $\left(x-\mu_X\right)^2$	9.00	4.00			
Weighted sq. dev. $\left(x-\mu_X\right)^2 P_X(x)$	0.45	0.40			

Example random function and its PMF.

- **Step 1**: Calculate the expectation $\mu_X = \mathbb{E}X$ of X.
  In the above example, we have $\mu_X = \mathbb{E}X = +1.00$.
- **Step 2**: Calculate the deviation $x-\mu_X$, the square deviation $\left(x-\mu_X\right)^2$, and the weighted square deviation $\left(x-\mu_X\right)^2 P_X(x)$ for every $x \in \mathscr{X}$.
- **Step 3**:

Example of variance calculation

x	-2	-1	0	$+1$	$+2$
Probability mass $P_X(x)$	0.05	0.10	0.20	0.10	0.55
Deviation $x - \mu_x$	-3.00	-2.00	-1.00	± 0.00	$+1.00$
Square deviation $\left(x - \mu_X\right)^2$	9.00	4.00	1.00	0.00	1.00
Weighted sq. dev. $\left(x - \mu_X\right)^2 P_X(x)$	0.45	0.40	0.20	0.00	0.55

Example random function and its PMF.

- **Step 1**: Calculate the expectation $\mu_X = \mathbb{E}X$ of X.
  In the above example, we have $\mu_X = \mathbb{E}X = +1.00$.
- **Step 2**: Calculate the deviation $x - \mu_X$, the square deviation $\left(x - \mu_X\right)^2$, and the weighted square deviation $\left(x - \mu_X\right)^2 P_X(x)$ for every $x \in \mathscr{X}$.
- **Step 3**:

Example of variance calculation

x	-2	-1	0	$+1$	$+2$
Probability mass $P_X(x)$	0.05	0.10	0.20	0.10	0.55
Deviation $x - \mu_x$	-3.00	-2.00	-1.00	± 0.00	$+1.00$
Square deviation $(x - \mu_X)^2$	9.00	4.00	1.00	0.00	1.00
Weighted sq. dev. $(x - \mu_X)^2 P_X(x)$	0.45	0.40	0.20	0.00	0.55

Example random function and its PMF.

- **Step 1**: Calculate the expectation $\mu_X = \mathbb{E}X$ of X.
  In the above example, we have $\mu_X = \mathbb{E}X = +1.00$.
- **Step 2**: Calculate the deviation $x - \mu_X$, the square deviation $(x - \mu_X)^2$, and the weighted square deviation $(x - \mu_X)^2 P_X(x)$ for every $x \in \mathscr{X}$.
- **Step 3**: Take the sum $\sum_{x \in \mathscr{X}} (x - \mu_X)^2 P_X(x)$.
  In the above example, we have
  $\mathbb{V}[X] = \sum_{x \in \mathscr{X}} (x - \mu_X)^2 P_X(x) = 0.45 + 0.40 + 0.20 + 0.00 + 0.55$

Example of variance calculation

x	-2	-1	0	$+1$	$+2$
Probability mass $P_X(x)$	0.05	0.10	0.20	0.10	0.55
Deviation $x - \mu_x$	-3.00	-2.00	-1.00	± 0.00	$+1.00$
Square deviation $\left(x - \mu_X\right)^2$	9.00	4.00	1.00	0.00	1.00
Weighted sq. dev. $\left(x - \mu_X\right)^2 P_X(x)$	0.45	0.40	0.20	0.00	0.55

Example random function and its PMF.

- **Step 1**: Calculate the expectation $\mu_X = \mathbb{E} X$ of X.
  In the above example, we have $\mu_X = \mathbb{E} X = +1.00$.
- **Step 2**: Calculate the deviation $x - \mu_X$, the square deviation $\left(x - \mu_X\right)^2$, and the weighted square deviation $\left(x - \mu_X\right)^2 P_X(x)$ for every $x \in \mathscr{X}$.
- **Step 3**: Take the sum $\sum_{x \in \mathscr{X}} \left(x - \mu_X\right)^2 P_X(x)$.
  In the above example, we have
  $\mathbb{V}[X] = \sum_{x \in \mathscr{X}} \left(x - \mu_X\right)^2 P_X(x) = 0.45 + 0.40 + 0.20 + 0.00 + 0.55 = 1.60.$

Another formula of the variance

The following formula is also useful.

Theorem

Let X be a discrete random variable taking values in $\mathscr{X} \subset \mathbb{R}$. Also, suppose that $\mu := \mathbb{E}X$ and $P_X : \mathscr{X} \to [0,1]$ are its expectation and PMF, respectively.
The variance $\mathbb{V}[X]$ is given by

$$\mathbb{V}[X] = \mathbb{E}X^2 - \mu^2 = \sum_{x \in \mathscr{X}} x^2 P_X(x) - \left(\sum_{x \in \mathscr{X}} x P_X(x) \right)^2. \tag{14}$$

Proof.

$$\mathbb{V}[X] = \mathbb{E}\left[(X-\mu)^2 \right] = \mathbb{E}\left[X^2 - 2\mu X + \mu^2 \right] = \mathbb{E}X^2 - 2\mu \cdot \mu + \mu^2 = \mathbb{E}X^2 - \mu^2. \tag{15}$$

$\square$

The variance of a linear function

Theorem

Let X be a random variable and $a, b \in \mathbb{R}$ be real numbers. Then we have that

$$\mathbb{V}[aX + b] = a^2 \mathbb{V}[X]. \tag{16}$$

In particular, the variance does not depend on b.

Example of calculating the variance of a linear function

Example

Suppose that X is a random variable whose PMF P_X is given by the following table.

x	-2	-1	0	$+1$	$+2$
$P_X(x)$	0.05	0.10	0.20	0.10	0.55

Example random function and its PMF.

The variance is given by $\mathbb{V}[X] = 1.60$.
Let's consider the random function given by $-3X + 5$ and its variance.
According to the formula, $\mathbb{V}[-3X+5] = (-3)^2\mathbb{V}[X] = (-3)^2 \cdot 1.60 = 14.40$.
Note that we did not use the PMF of $-3X + 5$ to calculate $\mathbb{V}[-3X + 5]$.

Standard deviation

Variance's interpretation is somewhat tricky since its effect against scaling 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

$$\sigma[X] := \sqrt{\mathbb{V}[X]}. \tag{17}$$

Example of the standard deviation calculation

Example

Suppose that X is a random variable whose PMF P_X is given by the following table.

x	-2	-1	0	$+1$	$+2$
$P_X(x)$	0.05	0.10	0.20	0.10	0.55

Example random function and its PMF.

The variance is given by $\mathbb{V}[X] = 1.60$.
Hence, the standard deviation $\sigma[X]$ is given by $\sigma[X] = \sqrt{\mathbb{V}[X]} = \sqrt{1.60} = 1.2649\ldots$

The standard deviation of a linear function

Theorem

If f is a linear function, i.e., if $f(x) = ax + b$, where $a, b \in \mathbb{R}$, then we have that

$$\sigma[f(X)] = \sigma[aX + b] = |a|\sigma[X]. \tag{18}$$

In particular, the standard deviation does not depend on b.

Hence, as we expected, the standard deviation of $10X$ is 10 times as large as that of X. In this sense, the standard deviation is "linear."

Note that the standard deviation is always non-negative. In particular, $\sigma[-10X]$ equals $10\sigma[X]$, but not $-10\sigma[X]$. This is an expected behavior since we originally wanted to measure the variability, which does not change even if we flip the sign.

Outline

Exercise (Frequency)

Suppose that we have $m = 20$ students and consider their results in an exam. For $x \in \mathscr{X} = \{0, 1, 2, 3, 4, 5\}$, we denote the number of the students who got a score x by m_x. Suppose that m_x is given by the following table.

Score x	0	1	2	3	4	5
# students m_x	3	2	3	5	6	1

Exam results.

Let X be the score of the student sampled uniform-randomly from the 20 students. Find the PMF of X.

Exercise (Expectation and median)

Suppose that X is a random variable whose PMF P_X is given by the following table.

x	-2	-1	0	$+1$	$+2$
$P_X(x)$	0.05	0.10	0.20	0.10	0.55

Example random function and its PMF.

(i) Find the value of the expectation $\mathbb{E}X$.

(ii) Find the value of the median of X.

Exercise (Cumulative distribution function)

Suppose that X is a random variable whose PMF is given as follows.

x	0	1	2	3	4	5
$P_X(x) := \Pr(X = x)$	0.15	0.10	0.15	0.25	0.30	0.05

The PMF of a student exam result frequency

Find the cumulative distribution function of X and plot the graph.

Exercise (Variance)

Let X be a discrete random variable, whose PMF is given by the following table.

x	-2	-1	0	$+1$	$+2$
Probability mass $P_X(x)$	0.05	0.10	0.20	0.10	0.55

Example random function and its PMF.

1. Calculate the variance $\mathbb{V}[X]$.
2. Write down the PMF table of $X-2$.
3. Calculate the variance $\mathbb{V}[X-2]$.
4. Write down the PMF table of $5X$.
5. Calculate the variance $\mathbb{V}[5X]$.

Outline

Outline

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.

Learning outcomes

By the end of this topic, you should be able to:

- Explain why two probability mass functions are not sufficient to describe multiple random variables,
- Describe multiple random variables using the joint probability mass function and conditional probability mass function,
- Describe the relation between multiple random variables using covariance, correlation, and independence, and
- Explain the difference between covariance, correlation, independence, and causality.

Outline

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 $(x, y) \in \mathscr{X} \times \mathscr{Y}$. 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 $\mathscr{X}$ and $\mathscr{Y}$, respectively, where $\mathscr{X}, \mathscr{Y} \subset \mathbb{R}$. We define the ***joint probability mass function (joint PMF)*** $P_{X,Y} : \mathscr{X} \times \mathscr{Y} \to [0, 1]$ of the pair of random variables X, Y by

$$P_{X,Y}(x, y) := \Pr(X = x, Y = y). \tag{19}$$

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 = 0	1	2	3	$P_Y(y)$
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 $\mathscr{X}_1, \mathscr{X}_2, \ldots, \mathscr{X}_m \subset \mathbb{R}$, respectively. We define the ***joint probability mass function (joint PMF)*** $P_{X_1, X_2, \ldots, X_m} : \mathscr{X}_1 \times \mathscr{X}_2 \times \cdots \times \mathscr{X}_m \to [0, 1]$ of random variables $X_1, X_2, \ldots, X_m$ by

$$P_{X_1, X_2, \ldots, X_m}(x_1, x_2, \ldots, x_m) := \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m). \tag{20}$$

Outline

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

$$P_X(x) = \sum_{y \in \mathscr{Y}} P_{X,Y}(x, y), \quad P_Y(y) = \sum_{x \in \mathscr{X}} P_{X,Y}(x, y), \tag{21}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	
y	1	0.24	0.08	0.12	0.06	
	2	0.06	0.00	0.06	0.18	
$P_X(x)$						

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.

$$\begin{aligned} P_X(0) &= P_{X,Y}(0,0) + P_{X,Y}(0,1) + P_{X,Y}(0,2) \\ &= 0.10 + 0.24 + 0.06 \end{aligned} \tag{22}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	
y	1	0.24	0.08	0.12	0.06	
	2	0.06	0.00	0.06	0.18	
$P_X(x)$		**0.40**				

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.

$$\begin{aligned} P_X(0) &= P_{X,Y}(0,0) + P_{X,Y}(0,1) + P_{X,Y}(0,2) \\ &= 0.10 + 0.24 + 0.06 = \mathbf{0.40}. \end{aligned} \tag{22}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	
y	1	0.24	0.08	0.12	0.06	
	2	0.06	0.00	0.06	0.18	
$P_X(x)$		0.40				

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.

$$
\begin{aligned}
P_X(1) &= P_{X,Y}(1,0) + P_{X,Y}(1,1) + P_{X,Y}(1,2) \\
&= 0.02 + 0.08 + 0.00
\end{aligned}
\tag{22}
$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	
y	1	0.24	0.08	0.12	0.06	
	2	0.06	0.00	0.06	0.18	
$P_X(x)$		0.40	**0.10**			

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.

$$\begin{aligned} P_X(1) &= P_{X,Y}(1,0) + P_{X,Y}(1,1) + P_{X,Y}(1,2) \\ &= 0.02 + 0.08 + 0.00 = \mathbf{0.10}. \end{aligned} \tag{22}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	
y	1	0.24	0.08	0.12	0.06	
	2	0.06	0.00	0.06	0.18	
$P_X(x)$		0.40	0.10			

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.

$$\begin{aligned} P_X(2) &= P_{X,Y}(2,0) + P_{X,Y}(2,1) + P_{X,Y}(2,2) \\ &= 0.02 + 0.12 + 0.06 \end{aligned} \tag{22}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	
y	1	0.24	0.08	0.12	0.06	
	2	0.06	0.00	0.06	0.18	
$P_X(x)$		0.40	0.10	**0.20**		

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

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

$$\begin{aligned} P_X(2) &= P_{X,Y}(2,0) + P_{X,Y}(2,1) + P_{X,Y}(2,2) \\ &= 0.02 + 0.12 + 0.06 = \mathbf{0.20}. \end{aligned} \tag{22}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	
y	1	0.24	0.08	0.12	0.06	
	2	0.06	0.00	0.06	0.18	
$P_X(x)$		0.40	0.10	0.20		

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.

$$\begin{aligned} P_X(3) &= P_{X,Y}(3,0) + P_{X,Y}(3,1) + P_{X,Y}(3,2) \\ &= 0.06 + 0.06 + 0.18 \end{aligned} \tag{22}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	
y	1	0.24	0.08	0.12	0.06	
	2	0.06	0.00	0.06	0.18	
$P_X(x)$		0.40	0.10	0.20	**0.30**	

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.

$$\begin{aligned} P_X(3) &= P_{X,Y}(3,0) + P_{X,Y}(3,1) + P_{X,Y}(3,2) \\ &= 0.06 + 0.06 + 0.18 = \mathbf{0.30}. \end{aligned} \tag{22}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	
y	1	0.24	0.08	0.12	0.06	
	2	0.06	0.00	0.06	0.18	
$P_X(x)$		0.40	0.10	0.20	0.30	

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.

$$\begin{aligned} P_Y(0) &= P_{X,Y}(0,0) + P_{X,Y}(1,0) + P_{X,Y}(2,0) + P_{X,Y}(3,0) \\ &= 0.10 + 0.02 + 0.02 + 0.06 \end{aligned} \tag{22}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	**0.20**
y	1	0.24	0.08	0.12	0.06	
	2	0.06	0.00	0.06	0.18	
$P_X(x)$		0.40	0.10	0.20	0.30	

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

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

$$\begin{aligned} P_Y(0) &= P_{X,Y}(0,0) + P_{X,Y}(1,0) + P_{X,Y}(2,0) + P_{X,Y}(3,0) \\ &= 0.10 + 0.02 + 0.02 + 0.06 = \mathbf{0.20}. \end{aligned} \tag{22}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	0.20
y	1	0.24	0.08	0.12	0.06	
	2	0.06	0.00	0.06	0.18	
$P_X(x)$		0.40	0.10	0.20	0.30	

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.

$$\begin{aligned} P_Y(1) &= P_{X,Y}(0,1) + P_{X,Y}(1,1) + P_{X,Y}(2,1) + P_{X,Y}(3,1) \\ &= 0.24 + 0.08 + 0.12 + 0.06 \end{aligned} \tag{22}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	0.20
y	1	0.24	0.08	0.12	0.06	**0.50**
	2	0.06	0.00	0.06	0.18	
$P_X(x)$		0.40	0.10	0.20	0.30	

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.

$$\begin{aligned} P_Y(1) &= P_{X,Y}(0,1) + P_{X,Y}(1,1) + P_{X,Y}(2,1) + P_{X,Y}(3,1) \\ &= 0.24 + 0.08 + 0.12 + 0.06 = \mathbf{0.50}. \end{aligned} \tag{22}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	0.20
y	1	0.24	0.08	0.12	0.06	0.50
	2	0.06	0.00	0.06	0.18	
$P_X(x)$		0.40	0.10	0.20	0.30	

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.

$$\begin{aligned} P_Y(2) &= P_{X,Y}(0,2) + P_{X,Y}(1,2) + P_{X,Y}(2,2) + P_{X,Y}(3,2) \\ &= 0.06 + 0.00 + 0.06 + 0.18 \end{aligned} \tag{22}$$

Marginal distribution example

		x				$P_Y(y)$
		0	1	2	3	
	0	0.10	0.02	0.02	0.06	0.20
y	1	0.24	0.08	0.12	0.06	0.50
	2	0.06	0.00	0.06	0.18	**0.30**
$P_X(x)$		0.40	0.10	0.20	0.30	

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.

$$\begin{aligned} P_Y(2) &= P_{X,Y}(0,2) + P_{X,Y}(1,2) + P_{X,Y}(2,2) + P_{X,Y}(3,2) \\ &= 0.06 + 0.00 + 0.06 + 0.18 = \mathbf{0.30}. \end{aligned} \tag{22}$$

Outline

Conditional distribution

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

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

The conditional distribution is the distribution of one RV when we know the value of the other RV.

The probability mass function (PMF) of the conditional distribution is called the ***conditional PMF***.

Conditional distribution example

Let X and Y be discrete RVs, and suppose that their joint PMF $P_{X,Y}$ and marginal PMFs P_X and P_Y are given by the following table.

		x = 0	1	2	3	$P_Y(y)$
y	0	0.10	0.02	0.02	0.06	0.20
	1	0.24	0.08	0.12	0.06	0.50
	2	0.06	0.00	0.06	0.18	0.30
$P_X(x)$		0.40	0.10	0.20	0.30	

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

Conditional distribution example

Let X and Y be discrete RVs, and suppose that their joint PMF $P_{X,Y}$ and marginal PMFs P_X and P_Y are given by the following table.

		x = 0	1	2	3	$P_Y(y)$
y	0	0.10	0.02	0.02	0.06	0.20
	1	0.24	0.08	0.12	0.06	0.50
	2	0.06	0.00	0.06	0.18	0.30
$P_X(x)$		0.40	0.10	0.20	0.30	

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

Suppose that we know that $Y = 2$. This information changes the distribution of X. For example, $X = 1$ no longer happens, so the probability of the event $X = 1$ is now zero.

So, for $x = 0, 1, 2, 3$, what is the probability of "$X = x$" when we know $Y = 2$? It is called the ***conditional probability*** of $X = x$ given $Y = 2$ and denoted by $P_{X|Y}(x|2)$.

Conditional probability calculation

	x				$P_Y(y)$
	0	1	2	3	
$P_{X,Y}(x,2)$	0.06	0.00	0.06	0.18	0.30

Joint PMF and conditional PMF

- If we know $Y = 2$, then the probability masses of $X = 0,1,2,3$ are proportional to the joint masses $P_{X,Y}(0,2), P_{X,Y}(1,2), P_{X,Y}(2,2), P_{X,Y}(3,2)$, shown above.
- The sum $P_{X|Y}(0|2) + P_{X|Y}(1|2) + P_{X|Y}(2|2) + P_{X|Y}(3|2)$ of the conditional probabilities must be 1 for them to be probabilities.

Hence, the conditional probability $P_{X|Y}(x|2)$ is each joint probability over the sum, i.e.,

$$P_{X|Y}(x|2) = \frac{P_{X,Y}(x,2)}{P_{X,Y}(0,2) + P_{X,Y}(1,2) + P_{X,Y}(2,2) + P_{X,Y}(3,2)} = \frac{P_{X,Y}(x,2)}{P_Y(2)}. \quad (23)$$

Conditional probability calculation example

	x = 0	1	2	3	$P_Y(y)$
$P_{X,Y}(x,2)$	0.06	0.00	0.06	0.18	0.30
$P_{X\|Y}(x\|y)$					

Joint PMF and conditional PMF

For example,

Conditional probability calculation example

	x 0	1	2	3	$P_Y(y)$
$P_{X,Y}(x,2)$	0.06	0.00	0.06	0.18	0.30
$P_{X\|Y}(x\|y)$	?				

Joint PMF and conditional PMF

For example,

$$P_{X|Y}(0|2) = \frac{P_{X,Y}(0,2)}{P_Y(2)} \tag{23}$$

Conditional probability calculation example

	x = 0	1	2	3	$P_Y(y)$
$P_{X,Y}(x,2)$	0.06	0.00	0.06	0.18	0.30
$P_{X\|Y}(x\|y)$	**0.20**				

Joint PMF and conditional PMF

For example,

$$P_{X|Y}(0|2) = \frac{P_{X,Y}(0,2)}{P_Y(2)} = \frac{\mathbf{0.06}}{\mathbf{0.30}} = \mathbf{0.20} \tag{23}$$

Conditional probability calculation example

	x				$P_Y(y)$
	0	1	2	3	
$P_{X,Y}(x,2)$	0.06	0.00	0.06	0.18	0.30
$P_{X\|Y}(x\|y)$	0.20	?			

Joint PMF and conditional PMF

For example,

$$P_{X|Y}(1|2) = \frac{P_{X,Y}(1,2)}{P_Y(2)} \tag{23}$$

Conditional probability calculation example

	x = 0	1	2	3	$P_Y(y)$
$P_{X,Y}(x,2)$	0.06	0.00	0.06	0.18	0.30
$P_{X\|Y}(x\|y)$	0.20	**0.00**			

Joint PMF and conditional PMF

For example,

$$P_{X|Y}(1|2) = \frac{P_{X,Y}(1,2)}{P_Y(2)} = \frac{\mathbf{0.00}}{\mathbf{0.30}} = \mathbf{0.00} \tag{23}$$

Conditional probability calculation example

	x				$P_Y(y)$
	0	1	2	3	
$P_{X,Y}(x,2)$	0.06	0.00	0.06	0.18	0.30
$P_{X\|Y}(x\|y)$	0.20	0.00	?		

Joint PMF and conditional PMF

For example,

$$P_{X|Y}(2|2) = \frac{P_{X,Y}(2,2)}{P_Y(2)} \tag{23}$$

Conditional probability calculation example

	x 0	1	2	3	$P_Y(y)$
$P_{X,Y}(x,2)$	0.06	0.00	0.06	0.18	0.30
$P_{X\|Y}(x\|y)$	0.20	0.00	**0.20**		

Joint PMF and conditional PMF

For example,

$$P_{X|Y}(2|2) = \frac{P_{X,Y}(2,2)}{P_Y(2)} = \frac{\mathbf{0.06}}{\mathbf{0.30}} = \mathbf{0.20} \tag{23}$$

Conditional probability calculation example

	x				$P_Y(y)$
	0	1	2	3	
$P_{X,Y}(x,2)$	0.06	0.00	0.06	0.18	0.30
$P_{X\|Y}(x\|y)$	0.20	0.00	0.20	?	

Joint PMF and conditional PMF

For example,

$$P_{X|Y}(3|2) = \frac{P_{X,Y}(3,2)}{P_Y(2)} \tag{23}$$

Conditional probability calculation example

	x				$P_Y(y)$
	0	1	2	3	
$P_{X,Y}(x,2)$	0.06	0.00	0.06	0.18	0.30
$P_{X\|Y}(x\|y)$	0.20	0.00	0.20	**0.60**	

Joint PMF and conditional PMF

For example,

$$P_{X|Y}(3|2) = \frac{P_{X,Y}(3,2)}{P_Y(2)} = \frac{\mathbf{0.18}}{\mathbf{0.30}} = \mathbf{0.60} \tag{23}$$

Conditional probability calculation example

	x = 0	1	2	3	$P_Y(y)$
$P_{X,Y}(x,2)$	0.06	0.00	0.06	0.18	0.30
$P_{X\|Y}(x\|y)$	0.20	0.00	0.20	0.60	
$P_X(x)$	0.40	0.10	0.20	0.30	

Joint PMF and conditional PMF

You can see that

- The conditional probabilities are different from the marginal probabilities.
- The sum $P_{X|Y}(0|2) + P_{X|Y}(1|2) + P_{X|Y}(2|2) + P_{X|Y}(3|2)$ of the conditional probabilities is one.

We call the function $P_{X|Y}$ the ***conditional PMF*** of X given Y.

Definition of the conditional PMF

Definition

Let X and Y be discrete random variables, whose supports are $\mathscr{X}$ and $\mathscr{Y}$, respectively. In other words, for any $x \in \mathscr{X}$ and $y \in \mathscr{Y}$, $P_X(x) > 0$ and $P_Y(y) > 0$ holds, where P_X and P_Y are the marginal PMFs of X and Y, respectively.
Let $P_{X,Y}$ be the joint PMF of X and Y.
We define the conditional PMF $P_{X|Y}$ by

$$P_{X|Y}(x|y) := \frac{P_{X,Y}(x,y)}{P_Y(y)}. \tag{23}$$

Likewise, we define the conditional PMF $P_{Y|X}$ by

$$P_{Y|X}(y|x) := \frac{P_{X,Y}(x,y)}{P_X(x)}. \tag{24}$$

Note: The conditional probability is not commutable.

Note that $P_{X|Y}(x|y) \neq P_{Y|X}(y|x)$ in general.

In this sense, the conditional probability is **NOT commutable**.

Outline

Independence of random variables

Suppose that the conditional PMF always equals 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[5] holds, we say that X and Y are independent.

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

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

Example of independent random variables

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

		x = 0	1	2	3	$P_Y(y)$
	0	0.08	0.02	0.04	0.06	0.20
y	1	0.24	0.08	0.12	0.06	0.50
	2	0.06	0.00	0.06	0.18	0.30
$P_X(x)$		0.40	0.10	0.20	0.30	

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

We can confirm that X and Y are mutually independent by checking that $P_{X,Y}(x,y) = P_X(x)P_Y(y)$ holds for every $x \in \mathscr{X} = \{0,1,2,3\}$ and $y \in \mathscr{Y} = \{0,1,2\}$.

Example of independent random variables

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

		x 0	1	2	3	$P_Y(y)$
	0	**0.08**	0.02	0.04	0.06	0.20
y	1	0.24	0.08	0.12	0.06	0.50
	2	0.06	0.00	0.06	0.18	0.30
$P_X(x)$		0.40	0.10	0.20	0.30	

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

For example, $P_{X,Y}(0,0) = 0.08$, which equals to $P_X(0)P_Y(0) = 0.40 \times 0.20$.

Example of independent random variables

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

		x 0	1	2	3	$P_Y(y)$
	0	0.08	0.02	0.04	0.06	0.20
y	1	0.24	0.08	**0.12**	0.06	0.50
	2	0.06	0.00	0.06	0.18	0.30
$P_X(x)$		0.40	0.10	0.20	0.30	

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

For example, $P_{X,Y}(2, 1) = 0.10$, which equals to $P_X(2)P_Y(1) = 0.20 \times 0.50$.

Conditional probability calculation

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

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}. \tag{25}$$

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.

Outline

Summary statistics for multiple RVs to show the relation

When we have multiple variables, we can calculate summary statistics for each of the variables. However, they do not give us information about the relation between multiple variables.

There are some statistics to show the relation between two RVs.

One principal question about the relation between two random variables X and Y is: “Do the RVs tend to take (relatively) large values simultaneously?”

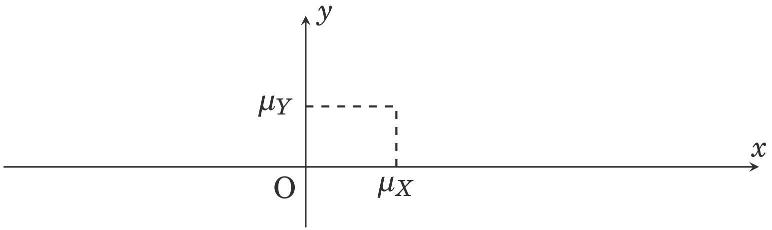
If X is easily observable and Y is the value of some product in the near future, then the information about the above relation financially benefits us.

The idea of covariance

The question is “Do the RVs tend to take (relatively) large values simultaneously?”

To answer the question, we consider the product of $X - \mu_X$ and $Y - \mu_Y$, where $\mu_X := \mathbb{E}X$ and $\mu_Y := \mathbb{E}X$ are the expectations of X and Y, respectively.

The value $X - \mu_X$ is positive if X takes a relatively large value and negative if X takes a relatively small value.

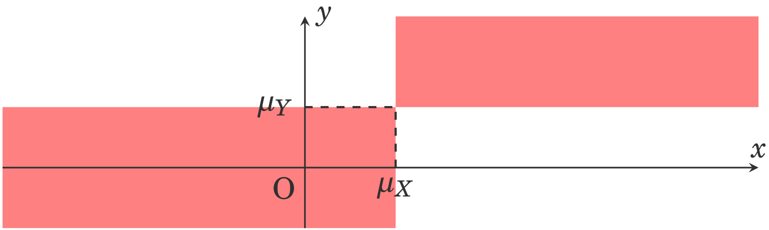


The idea of covariance

The question is "Do the RVs tend to take (relatively) large values simultaneously?"

To answer the question, we consider the product of $X - \mu_X$ and $Y - \mu_Y$, where $\mu_X := \mathbb{E}X$ and $\mu_Y := \mathbb{E}X$ are the expectations of X and Y, respectively.

If X and Y tend to take large values simultaneously and small values simultaneously as well, then the product $(X - \mu_X)(Y - \mu_Y)$ tends to be positive.



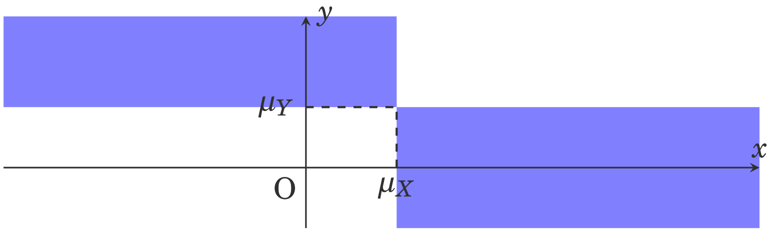
The area where $(X - \mu_X)(Y - \mu_Y)$ takes a positive value.

The idea of covariance

The question is "Do the RVs tend to take (relatively) large values simultaneously?"

To answer the question, we consider the product of $X - \mu_X$ and $Y - \mu_Y$, where $\mu_X := \mathbb{E}X$ and $\mu_Y := \mathbb{E}X$ are the expectations of X and Y, respectively.

Conversely, if one tends to be small when the other is large, then the product $(X - \mu_X)(Y - \mu_Y)$ tends to be negative.



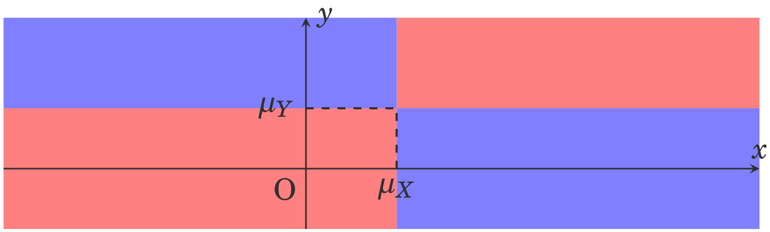
The area where $(X - \mu_X)(Y - \mu_Y)$ takes a negative value.

The idea of covariance

The question is “Do the RVs tend to take (relatively) large values simultaneously?”

To answer the question, we consider the product of $X - \mu_X$ and $Y - \mu_Y$, where $\mu_X := \mathbb{E}X$ and $\mu_Y := \mathbb{E}X$ are the expectations of X and Y, respectively.

Hence, we are interested in the value of $(X - \mu_X)(Y - \mu_Y)$. This is the basic idea of ***covariance***. But what is $(X - \mu_X)(Y - \mu_Y)$?



A function of multiple RVs

We say that the variable $(X - \mu_X)(Y - \mu_Y)$ is a function of RVs X and Y since it depends on the RVs X and Y.

We remark that $(X - \mu_X)(Y - \mu_Y)$ is a random variable. In particular, it is a discrete RV since X and Y are discrete RVs. Since it is a random variable, we can define its expectation $\mathbb{E}(X - \mu_X)(Y - \mu_Y)$.

Let's discuss the general function of multiple RVs and define its expectations.

A function of multiple RVs and its expectation

Theorem

Suppose that X and Y are random variables and $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ are a real-valued function taking two real values as an input. Then, $f(X,Y)$ is a random variable.
In particular, suppose that X and Y are discrete RVs, their supports are $\mathscr{X}$ and $\mathscr{Y}$, respectively, and their joint PMF is $P_{X,Y}$. Then, $f(X,Y)$ is also a discrete RV and

- *The support of $f(X,Y)$ is $\{f(x,y) | x \in \mathscr{X}, y \in \mathscr{Y}\}$,*
- *The PMF $P_{f(X,Y)}$ is given by*

$$P_{f(X,Y)}(z) = \sum_{(x,y)\in\{(x',y')|f(x',y')=z\}} P_{X,Y}(x,y), \tag{26}$$

- *The expectation $\mathbb{E} f(X,Y)$ is given by*

$$\mathbb{E} f(X,Y) = \sum_{(x,y)\in\mathscr{X}\times\mathscr{Y}} f(x,y) P_{X,Y}(x,y). \tag{27}$$

The linearity of the expectation: the multi-variable case

From the linearity of the expectation operator $\mathbb{E}$, the following holds.

Theorem (The linearity of the expectation)

Let X, Y be random variables, $a, b \in \mathbb{R}$ be real numbers, and $f, g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be real-valued functions taking two real variables as an input. Then, we have that

$$\mathbb{E}[af(X,Y) + bg(X,Y)] = a\mathbb{E}f(X,Y) + b\mathbb{E}g(X,Y). \tag{28}$$

The above theorem provides us with the formula for the expectation calculation of a linear function of multiple variables.

Corollary

Let X, Y be random variables and $a, b, c \in \mathbb{R}$ be real numbers. Then, we have that

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}X + b\mathbb{E}Y + c. \tag{29}$$

Definition of the covariance

Now, we are ready to define the ***covariance***. Recall that the idea of covariance is to evaluate the behavior of $(X - \mu_X)(Y - \mu_Y)$. In fact, the covariance is nothing but the expectation of $(X - \mu_X)(Y - \mu_Y)$.

Definition (Covariance)

Let X and Y be RVs and $\mu_X := \mathbb{E}X$ and $\mu_Y := \mathbb{E}Y$ be their expectations. We define the ***covariance*** $\mathrm{Cov}(X,Y) \in \mathbb{R}$ between the two random variables X and Y by

$$\mathrm{Cov}(X,Y) := \mathbb{E}\big[(X - \mu_X)(Y - \mu_Y)\big]. \tag{30}$$

Note that the covariance is symmetric, i.e., $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

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 tends to take a relatively small value.

Formulae to calculate the covariance

We provide the explicit calculation formula of the covariance.

Theorem

Suppose that X and Y are discrete RVs, their supports are $\mathscr{X}$ and $\mathscr{Y}$, respectively, and their joint PMF is $P_{X,Y}$.
Then, the covariance $\mathrm{Cov}(X,Y) \in \mathbb{R}$ between the two random variables X and Y is given by

$$\mathrm{Cov}(X,Y) = \sum_{(x,y)\in\mathscr{X}\times\mathscr{Y}} (x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y). \tag{31}$$

Example of covariance calculation

Example

		x: 0	x: +1	$P_Y(y)$
	0	0.25	0.00	0.25
y	+1	0.25	0.25	0.50
	+2	0.00	0.25	0.25
$P_X(x)$		0.50	0.50	

The joint PMF $P_{X,Y}$. The RVs X and Y have a positive covariance.

We can calculate the covariance $\mathrm{Cov}(X,Y)$ of RVs X,Y from its joint PMF $P_{X,Y}$.

-

Example of covariance calculation

Example

		x = 0	x = +1	$P_Y(y)$
	0	0.25	0.00	0.25
y	+1	0.25	0.25	0.50
	+2	0.00	0.25	0.25
$P_X(x)$		0.50	0.50	

The joint PMF $P_{X,Y}$. The RVs X and Y have a positive covariance.

We can calculate the covariance $\mathrm{Cov}(X,Y)$ of RVs X,Y from its joint PMF $P_{X,Y}$.

- **Step 1**: Calculate the expectations $\mu_X = \mathbb{E}X$ and $\mu_Y = \mathbb{E}Y$. Then memorize the value $x - \mu_X$ for all $x \in \mathscr{X}$ and $y - \mu_Y$ for all $y \in \mathscr{Y}$.

Example of covariance calculation

Example

		x: 0	$+1$	$P_Y(y)$
	0	0.25	0.00	0.25
y	$+1$	0.25	0.25	0.50
	$+2$	0.00	0.25	0.25
$P_X(x)$		0.50	0.50	

The joint PMF $P_{X,Y}$. The RVs X and Y have a positive covariance.

We can calculate the covariance $\mathrm{Cov}(X,Y)$ of RVs X,Y from its joint PMF $P_{X,Y}$.

- **Step 1**: Calculate the expectations $\mu_X = \mathbb{E}X$ and $\mu_Y = \mathbb{E}Y$. Then memorize the value $x - \mu_X$ for all $x \in \mathscr{X}$ and $y - \mu_Y$ for all $y \in \mathscr{Y}$.
  In the above example, we have $\mu_X = \mathbb{E}X = +0.50$ and $\mu_Y = \mathbb{E}Y = +1.00$.

Example of covariance calculation

Example

		$x-\mu_X$		$P_Y(y)$
		-0.5	$+0.5$	
$y-\mu_Y$	-1	0.25	0.00	0.25
	0	0.25	0.25	0.50
	$+1$	0.00	0.25	0.25
$P_X(x)$		0.50	0.50	

The joint PMF $P_{X,Y}$. The RVs X and Y have a positive covariance.

We can calculate the covariance $\mathrm{Cov}(X,Y)$ of RVs X,Y from its joint PMF $P_{X,Y}$.

- **Step 1**: Calculate the expectations $\mu_X = \mathbb{E}X$ and $\mu_Y = \mathbb{E}Y$. Then memorize the value $x-\mu_X$ for all $x \in \mathscr{X}$ and $y-\mu_Y$ for all $y \in \mathscr{Y}$.
  In the above example, we have $\mu_X = \mathbb{E}X = +0.50$ and $\mu_Y = \mathbb{E}Y = +1.00$.

Example of covariance calculation

Example

		$x-\mu_X$		$P_Y(y)$
		-0.5	$+0.5$	
	-1	0.25	0.00	0.25
$y-\mu_Y$	0	0.25	0.25	0.50
	$+1$	0.00	0.25	0.25
$P_X(x)$		0.50	0.50	

The joint PMF $P_{X,Y}$. The RVs X and Y have a positive covariance.

We can calculate the covariance $\mathrm{Cov}(X,Y)$ of RVs X,Y from its joint PMF $P_{X,Y}$.

- **Step 2**: Calculate the weighted product of the deviations $(x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$ for every $(x,y) \in \mathscr{X} \times \mathscr{Y}$ and take the sum.
  In the above example, we have
  $\mathrm{Cov}(X,Y) = \sum_{(x,y)\in\mathscr{X}\times\mathscr{Y}} (x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$
  $= (\mathbf{-0.5}) \cdot (\mathbf{-1}) \cdot \mathbf{0.25}$

Example of covariance calculation

Example

		$x-\mu_X$		$P_Y(y)$
		-0.5	$+0.5$	
	-1	0.25	0.00	0.25
$y-\mu_Y$	0	0.25	0.25	0.50
	$+1$	0.00	0.25	0.25
$P_X(x)$		0.50	0.50	

The joint PMF $P_{X,Y}$. The RVs X and Y have a positive covariance.

We can calculate the covariance $\mathrm{Cov}(X,Y)$ of RVs X,Y from its joint PMF $P_{X,Y}$.

- **Step 2**: Calculate the weighted product of the deviations $(x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$ for every $(x,y)\in\mathscr{X}\times\mathscr{Y}$ and take the sum.
  In the above example, we have
  $\mathrm{Cov}(X,Y)=\sum_{(x,y)\in\mathscr{X}\times\mathscr{Y}}(x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$
  $=(-0.5)\cdot(-1)\cdot 0.25+(+\mathbf{0.5})\cdot(-\mathbf{1})\cdot\mathbf{0.00}$

Example of covariance calculation

Example

		$x-\mu_X$		$P_Y(y)$
		-0.5	$+0.5$	
	-1	0.25	0.00	0.25
$y-\mu_Y$	0	0.25	0.25	0.50
	$+1$	0.00	0.25	0.25
$P_X(x)$		0.50	0.50	

The joint PMF $P_{X,Y}$. The RVs X and Y have a positive covariance.

We can calculate the covariance $\mathrm{Cov}(X,Y)$ of RVs X,Y from its joint PMF $P_{X,Y}$.

- **Step 2**: Calculate the weighted product of the deviations $(x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$ for every $(x,y)\in\mathscr{X}\times\mathscr{Y}$ and take the sum.
  In the above example, we have
  $\mathrm{Cov}(X,Y)=\sum_{(x,y)\in\mathscr{X}\times\mathscr{Y}}(x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$
  $=(-0.5)\cdot(-1)\cdot 0.25+(+0.5)\cdot(-1)\cdot 0.00+\mathbf{(-0.5)\cdot 0\cdot 0.25}$

Example of covariance calculation

Example

		$x-\mu_X$		$P_Y(y)$
		-0.5	$+0.5$	
	-1	0.25	0.00	0.25
$y-\mu_Y$	0	0.25	0.25	0.50
	$+1$	0.00	0.25	0.25
$P_X(x)$		0.50	0.50	

The joint PMF $P_{X,Y}$. The RVs X and Y have a positive covariance.

We can calculate the covariance $\mathrm{Cov}(X,Y)$ of RVs X,Y from its joint PMF $P_{X,Y}$.

- **Step 2**: Calculate the weighted product of the deviations $(x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$ for every $(x,y)\in\mathscr{X}\times\mathscr{Y}$ and take the sum.
  In the above example, we have
  $\mathrm{Cov}(X,Y)=\sum_{(x,y)\in\mathscr{X}\times\mathscr{Y}}(x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$
  $=(-0.5)\cdot(-1)\cdot 0.25+(+0.5)\cdot(-1)\cdot 0.00+(-0.5)\cdot 0\cdot 0.25+\cdots+\mathbf{0.5\cdot 1\cdot 0.25}$

Example of covariance calculation

Example

		$x-\mu_X$: -0.5	$x-\mu_X$: $+0.5$	$P_Y(y)$
$y-\mu_Y$	-1	0.25	0.00	0.25
	0	0.25	0.25	0.50
	$+1$	0.00	0.25	0.25
$P_X(x)$		0.50	0.50	

The joint PMF $P_{X,Y}$. The RVs X and Y have a positive covariance.

We can calculate the covariance $\mathrm{Cov}(X,Y)$ of RVs X,Y from its joint PMF $P_{X,Y}$.

- **Step 2**: Calculate the weighted product of the deviations $(x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$ for every $(x,y) \in \mathscr{X} \times \mathscr{Y}$ and take the sum.
  In the above example, we have
  $\mathrm{Cov}(X,Y) = \sum_{(x,y)\in\mathscr{X}\times\mathscr{Y}} (x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$
  $= (-0.5)\cdot(-1)\cdot 0.25 + (+0.5)\cdot(-1)\cdot 0.00 + (-0.5)\cdot 0\cdot 0.25 + \cdots + 0.5\cdot 1\cdot 0.25 = 0.25.$

The variance is a special case of the covariance

The covariance between a random variable and itself is the variance of the random variable. In other words:

Theorem

$$\mathrm{Cov}(X, X) = \mathbb{V}[X]. \tag{32}$$

Covariance matrix

Definition

Let $X_1, X_2, \ldots, X_m$ be RVs. The $m \times m$ real matrix

$$\begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & (X_1, X_m) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & (X_2, X_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_m, X_1) & \mathrm{Cov}(X_m, X_2) & \cdots & (X_m, X_m) \end{bmatrix} \tag{33}$$

is called the ***covariance matrix*** of RVs $X_1, X_2, \ldots, X_m$.

Example of the covariance matrix

Let X and Y be random variables whose joint PMF $P_{X,Y}$ are given by the following table.

		x: 0	x: +1	$P_Y(y)$
y	0	0.25	0.00	0.25
	$+1$	0.25	0.25	0.50
	$+2$	0.00	0.25	0.25
$P_X(x)$		0.50	0.50	

The joint PMF $P_{X,Y}$. The RVs X and Y have a positive covariance.

In the above example, $\mathrm{Cov}(X,X) = \mathbb{V}[X] = 0.25$, $\mathrm{Cov}(Y,Y) = \mathbb{V}[Y] = 0.5$, and $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X) = 0.25$.

Hence, the covariance matrix is $\begin{bmatrix} \mathrm{Cov}(X,X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Cov}(Y,Y) \end{bmatrix} = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.5 \end{bmatrix}$.

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

$$\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X, Y). \tag{34}$$

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*** $\operatorname{corr}[X,Y]$ between X and Y is given by

$$\operatorname{corr}[X,Y] := \frac{\operatorname{Cov}[X,Y]}{\sigma[X]\sigma[Y]}. \tag{35}$$

The correlation coefficient is often denoted by ρ.

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

$$\operatorname{corr}[aX,bY] = \operatorname{corr}[X,Y]. \tag{36}$$

Example of the correlation coefficient

Let X and Y be random variables whose joint PMF $P_{X,Y}$ are given by the following table.

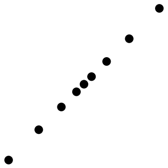
		x 0	+1	$P_Y(y)$
	0	0.25	0.00	0.25
y	+1	0.25	0.25	0.50
	+2	0.00	0.25	0.25
$P_X(x)$		0.50	0.50	

The joint PMF $P_{X,Y}$. The RVs X and Y have a positive covariance.

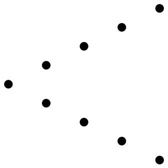
In the above example, $\mathrm{Cov}(X,X) = \mathbb{V}[X] = 0.25$, $\mathrm{Cov}(Y,Y) = \mathbb{V}[Y] = 0.5$, and $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X) = 0.25$.

Hence, the correlation coefficient between X and Y is $\mathrm{corr}(X,Y) = \frac{0.25}{\sqrt{0.25}\sqrt{0.5}} = \frac{1}{\sqrt{2}}$.

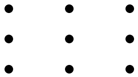
Correlation coefficient examples



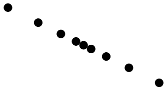
$\rho = 1.0$



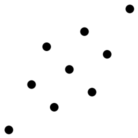
$\rho = 0.0$



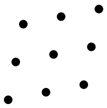
$\rho = 0.0$



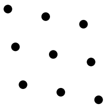
$\rho = -1.0$



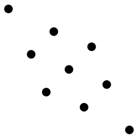
$\rho = 0.735$



$\rho = 0.385$



$\rho = -0.385$



$\rho = -0.735$

Independence implies no-correlation

Theorem

Let random variables X *and* Y *be mutually independent. Then the covariance* $\mathrm{Cov}(X, Y)$ *and the correlation* $\mathrm{corr}[X, Y]$ *are zero.*

Note: the converse of the above theorem is FALSE (see the next slide).

No correlation does NOT imply independence!

Example

Let X and Y be random variables whose joint PMF $P_{X,Y}$ are given by the following table.

		x: -1	x: $+1$	$P_Y(y)$
y	-1	0.0	0.25	0.25
	0	0.5	0.0	0.5
	$+1$	0.0	0.25	0.25
$P_X(x)$		0.5	0.5	

The joint PMF $P_{X,Y}$. The RVs X and Y are uncorrelated but mutually independent.

Then, the covariance $\mathrm{Cov}(X,Y)$ and the correlation $\mathrm{corr}[X,Y]$ are zero. However, X and Y are not independent. For example, $P_{X,Y}(-1,-1) \neq P_X(-1)P_Y(-1)$. The LHS is 0.0, while the RHS is $0.5 \times 0.25 = 0.125$.

No correlation does NOT imply independence!

Example

Let X and Y be random variables whose joint PMF $P_{X,Y}$ are given by the following table.

		x -1	x $+1$	$P_Y(y)$
y	-1	0.0	0.25	0.25
y	0	0.5	0.0	0.5
y	$+1$	0.0	0.25	0.25
$P_X(x)$		0.5	0.5	

The joint PMF $P_{X,Y}$. The RVs X and Y are uncorrelated but mutually independent.

Indeed, we cannot say Y increases as X increases since the expectation of Y is invariant when X. Hence, the correlation is zero. On the other hand, the variance of Y is 0 when $X = -1$ but it is non-zero if $X = +1$, hence X has some information about Y. These are intuitive explanations of zero correlation and non-independence of X and Y.

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 $\mathrm{corr}[X, Y] \neq 0$.

Outline

Exercise (Independent RVs (1))

Let X and Y be discrete RVs, whose supports are $\mathscr{X} = \{0, 1, 2, 3\}$ and $\mathscr{Y} = \{0, 1, 2\}$, respectively. Suppose that the joint PMF $P_{X,Y}$ is given by the following table.

		x = 0	1	2	3
y	0	0.08	0.02	0.04	0.06
	1	0.24	0.08	0.12	0.06
	2	0.06	0.00	0.06	0.18

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

(i) Find the values of the marginal PMFs $P_X(x)$ and $P_Y(y)$ for every $x \in \mathscr{X}$ and $y \in \mathscr{Y}$.
(ii) Judge whether X and Y are mutually independent or not.

Exercise (Independent RVs (2))

Let X and Y be discrete RVs, whose supports are $\mathscr{X} = \{0, 1, 2, 3\}$ and $\mathscr{Y} = \{0, 1, 2\}$, respectively. Suppose that the joint PMF $P_{X,Y}$ is given by the following table.

		x			
		0	1	2	3
	0	0.10	0.02	0.02	0.06
y	1	0.24	0.08	0.12	0.06
	2	0.06	0.00	0.06	0.18

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

(i) Find the values of the marginal PMFs $P_X(x)$ and $P_Y(y)$ for every $x \in \mathscr{X}$ and $y \in \mathscr{Y}$.
(ii) Judge whether X and Y are mutually independent or not.

Exercise (Joint PMF of independent RVs)

Let X and Y be mutually independent discrete RVs, whose supports are $\mathscr{X} = \{0, 1, 2, 3\}$ and $\mathscr{Y} = \{0, 1, 2\}$, respectively. Suppose that those marginal PMFs, P_X and P_Y, are given by the following tables.

	x			
	0	1	2	3
$P_X(x)$	0.40	0.10	0.20	0.30

	y		
	0	1	2
$P_Y(y)$	0.20	0.50	0.30

The marginal PMFs $P_X(x)$ and $P_Y(y)$.

Find the values of the Joint PMF $P_{X,Y}(x, y)$ for all $x \in \mathscr{X}$ and $y \in \mathscr{Y}$.

Exercise (Exercise: conditional PMF calculation)

Let X and Y be discrete RVs, whose supports are $\mathscr{X} = \{0, 1, 2, 3\}$ and $\mathscr{Y} = \{0, 1, 2\}$, respectively. Suppose that the joint PMF $P_{X,Y}$ is given by the following table.

		x			
		0	1	2	3
	0	0.08	0.02	0.04	0.06
y	1	0.24	0.08	0.12	0.06
	2	0.06	0.00	0.06	0.18

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

(i) Find the values of the marginal PMFs $P_X(x)$ and $P_Y(y)$ for every $x \in \mathscr{X}$ and $y \in \mathscr{Y}$.
(ii) Find the values of the conditional PMFs $P_{X|Y}(x|y)$ and $P_{X|Y}(x|x)$ for every $x \in \mathscr{X}$ and $y \in \mathscr{Y}$.

Exercise (Exercise: conditional PMF calculation)

Let X and Y be discrete RVs, whose supports are $\mathscr{X} = \{0, 1, 2, 3\}$ and $\mathscr{Y} = \{0, 1, 2\}$, respectively. Suppose that the joint PMF $P_{X,Y}$ is given by the following table.

		x			
		0	1	2	3
	0	0.10	0.02	0.02	0.06
y	1	0.24	0.08	0.12	0.06
	2	0.06	0.00	0.06	0.18

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

(i) Find the values of the marginal PMFs $P_X(x)$ and $P_Y(y)$ for every $x \in \mathscr{X}$ and $y \in \mathscr{Y}$.
(ii) Find the values of the conditional PMFs $P_{X|Y}(x|y)$ and $P_{X|Y}(x|x)$ for every $x \in \mathscr{X}$ and $y \in \mathscr{Y}$.

Exercise (Marginal distribution)

Suppose that the joint PMF of random variables X and Y is given by the following table. Find the marginal PMFs of X and Y.

		x = 0	1	2	3	$P_Y(y)$
	0	0.10	0.02	0.02	0.06	$P_Y(0) = ?$
y	1	0.24	0.08	0.12	0.06	$P_Y(1) = ?$
	2	0.06	0.00	0.06	0.18	$P_Y(2) = ?$
$P_X(x)$		$P_X(0) = ?$	$P_X(1) = ?$	$P_X(2) = ?$	$P_X(3) = ?$	

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

Outline

Outline

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 $\Pr(0 \le X \le 1) = 1$.

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, $\Pr(X \in A) \le N\epsilon$. However, there are a infinite number of real numbers in $[0, 1]$, so $\Pr(X \in [0, 1])$ is infinity. It contradicts $\Pr(X \in [0, 1]) = 1$. $\square$

A random variable may not have a PMF.

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

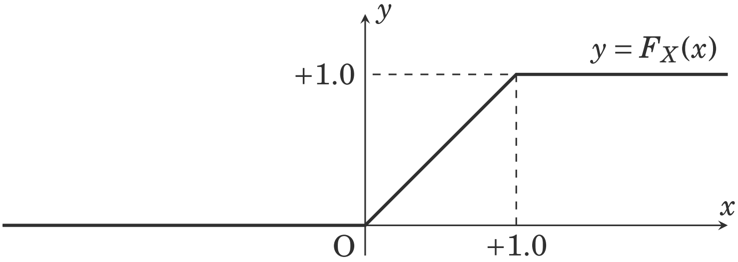
This random variable have nowhere probability mass, i.e., $\Pr(X = x) = 0$. for any $X \in \mathbb{R}$.

Other 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.

Fortunately, any univariate random variable has a cumulative distribution function (CDF)

CDF example

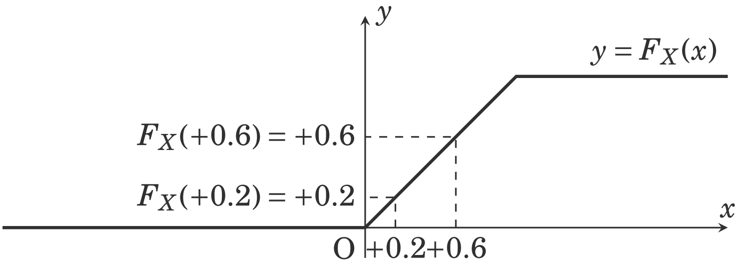
The CDF of a random variable X uniformly distributed in $[0,1]$ is:



$$F_X(x) = \begin{cases} 0 & \text{if } x \le 0, \\ x & \text{if } 0 \le x \le 1, \\ 1 & \text{if } x \ge 1. \end{cases} \tag{37}$$

CDF example

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

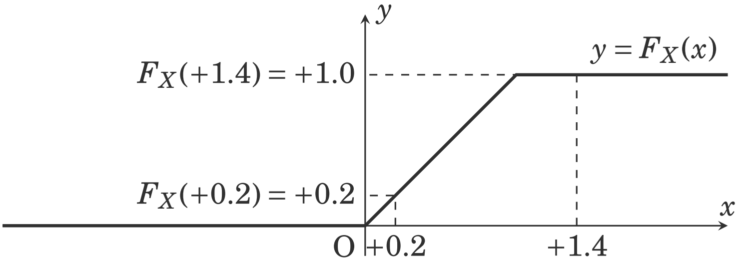


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned} \Pr(0.2 < X \leq 0.6) &= \Pr(X \leq 0.6) - \Pr(X \leq 0.2) \\ &= F_X(0.6) - F_X(0.2) \\ &= 0.6 - 0.2 = 0.4. \end{aligned} \tag{37}$$

CDF example

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

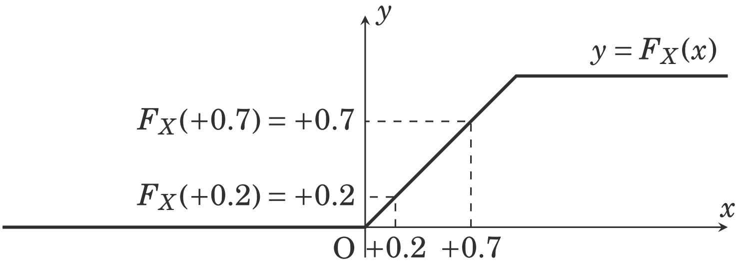


Using the CDF, we can calculate the probability of various events. For example,

$$
\begin{aligned}
\Pr(0.2 \le X \le 0.7) &= \Pr(X \le 0.7) - \lim_{x \nearrow 0.2} \Pr(x) \\
&= F_X(0.7) - \lim_{x \nearrow 0.2} F_X(x) \\
&= 0.7 - 0.2 = 0.5.
\end{aligned}
\tag{37}
$$

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,

Why are we not satisfied with the CDF?

- The CDF is not intuitive. At one glance, we do not know around which value the random variable tends to take a value.
- The CDF can be extremely complex even for a practically important distribution.

Although there exists no PMF for a continuous RV in general, we want to indicate which values the RV tends to take frequently as the PMF does for a discrete RV.

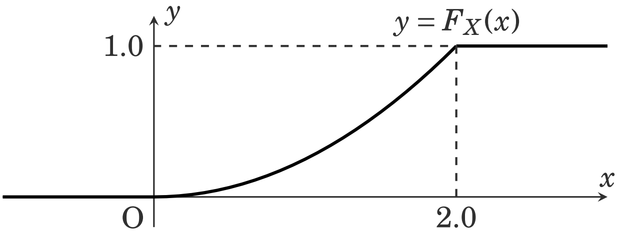
The ***probability density function (PDF)*** achieves this objective.

Learning outcomes

By the end of this topic, you should be able to:

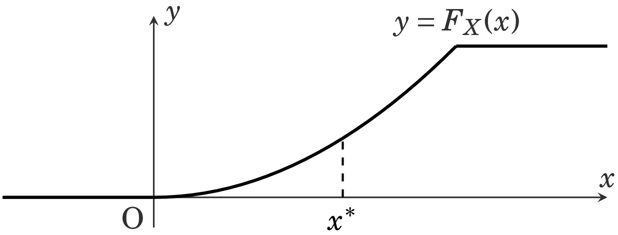
- Explain what a probability density function represents,
- Explain the relation between the probability density function and cumulative distribution function,
- Calculate the probability of an event using the integral and the probability density function, and
- Calculate summary statistics of continuous random variables.

CDF to the density.



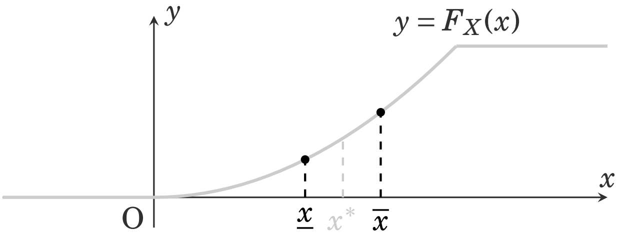
The CDF of a RV X.

CDF to the density.



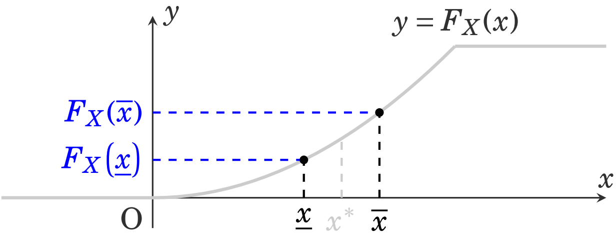
Suppose we want to know how frequently the RV X takes a value "around" x^*.

CDF to the density.



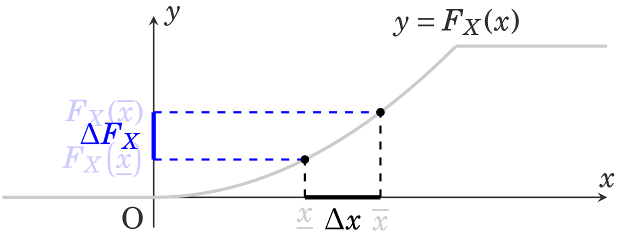
We consider an interval $[\underline{x}, \overline{x}]$ including x^*.

CDF to the density.



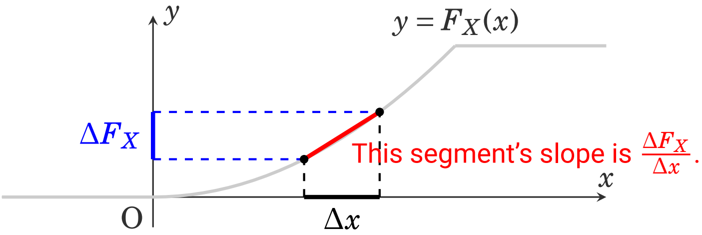
We find the probability $\Pr\left(X \in [\underline{x}, \overline{x}]\right)$, given by $F_X(\overline{x}) - F_X(\underline{x})$.

CDF to the density.



Define $\Delta x := \overline{x} - \underline{x}$ and $\Delta F_X := F_X(\overline{x}) - F_X(\underline{x}) = \Pr(X \in [\underline{x}, \overline{x}])$.

CDF to the density.

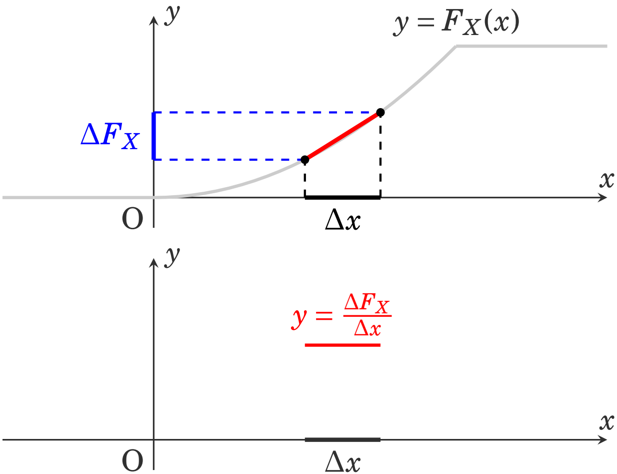


Define $\Delta x := \overline{x} - \underline{x}$ and $\Delta F_X := F_X(\overline{x}) - F_X(\underline{x}) = \Pr\left(X \in [\underline{x}, \overline{x}]\right)$.

We can say the RV X takes a value around x^* if the probability per length $\frac{\Delta F_X}{\Delta x}$, or the “density’

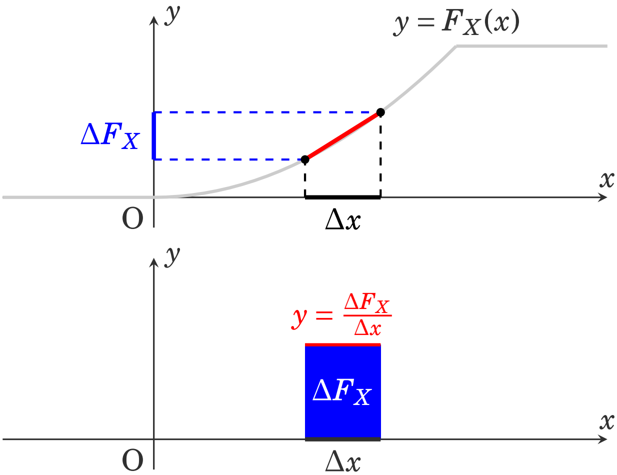
the probability per length $\frac{\Delta F_X}{\Delta x}$, or the “density”.

CDF to the density.



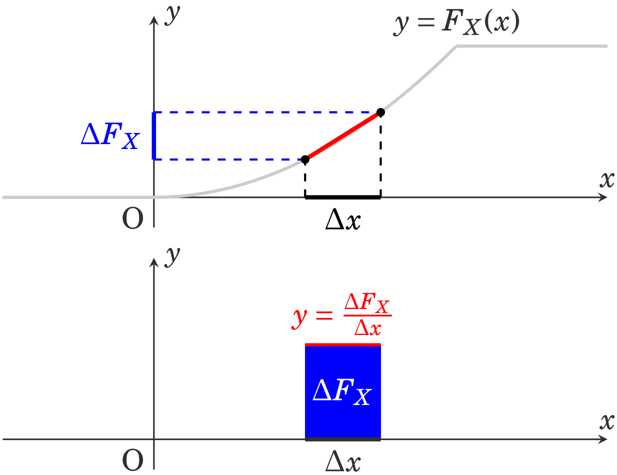
The graph plot of the probability per length $\frac{\Delta F_X}{\Delta x}$, or the “density”

CDF to the density.

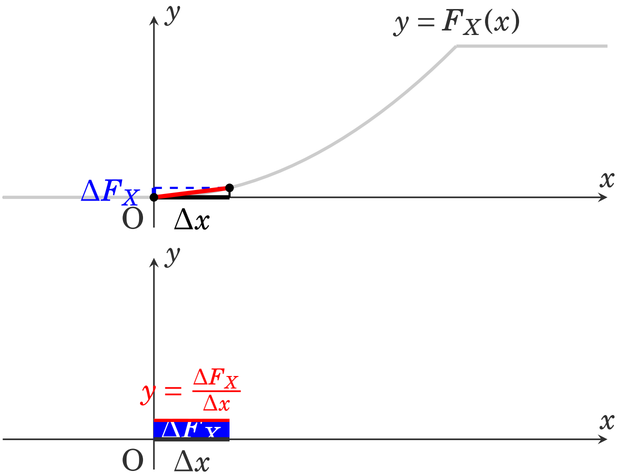


The area under the graph of $\frac{\Delta F_X}{\Delta x}$ is given by $\Delta x \cdot \frac{\Delta F_X}{\Delta x} = \Delta F_X$.

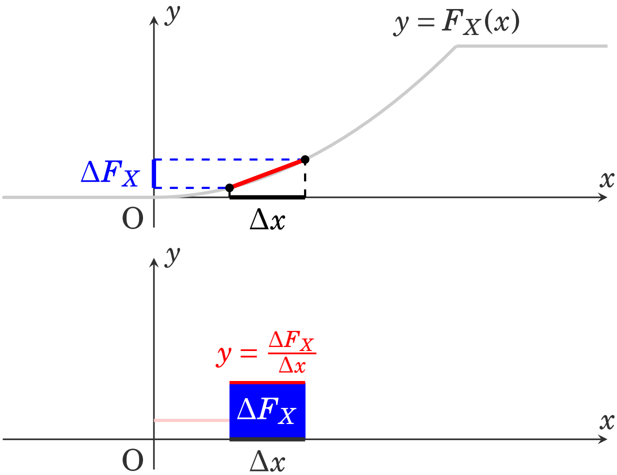
CDF to the density.



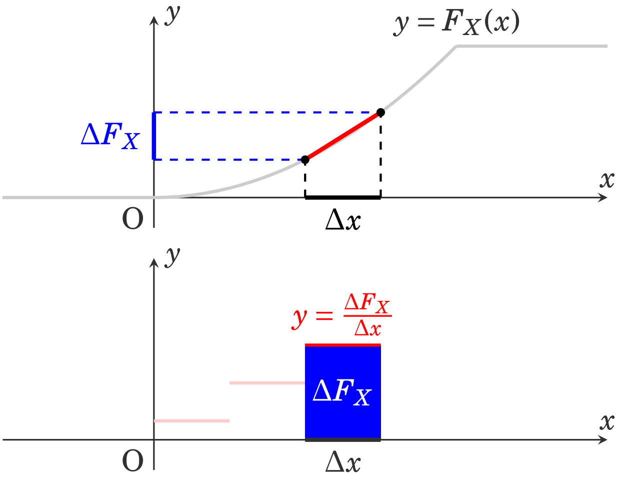
CDF to the density.



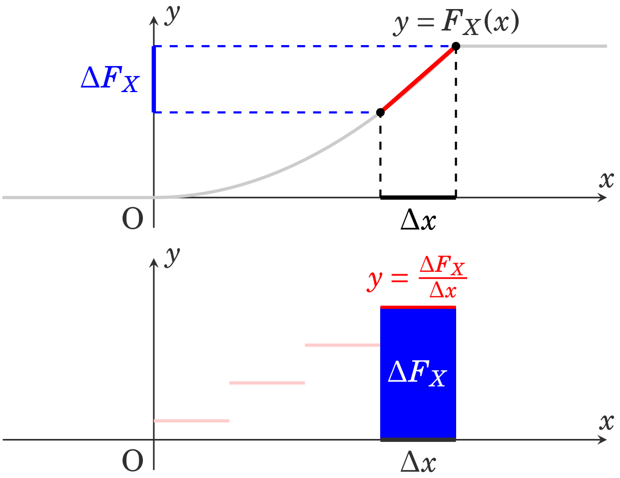
CDF to the density.



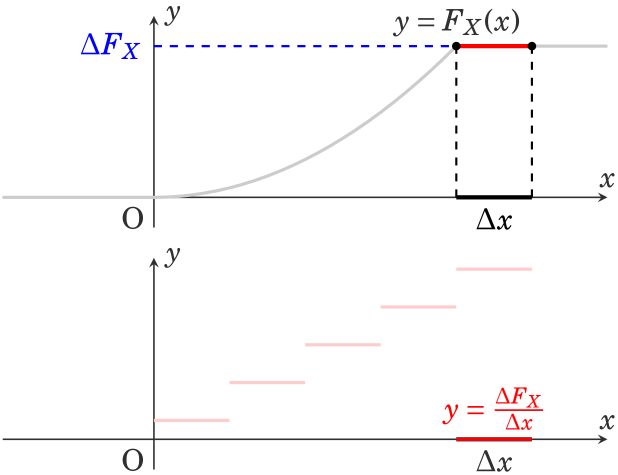
CDF to the density.



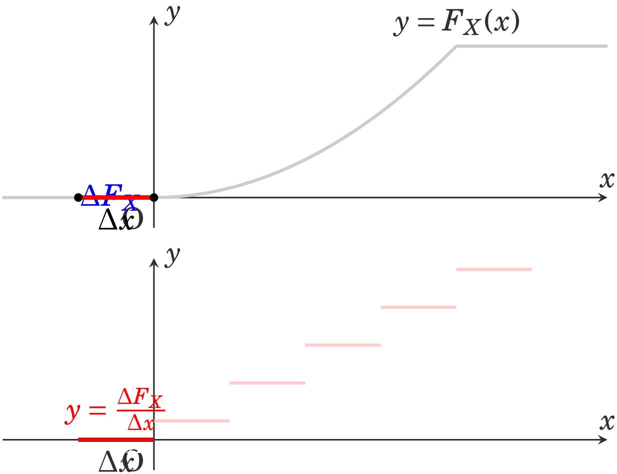
CDF to the density.



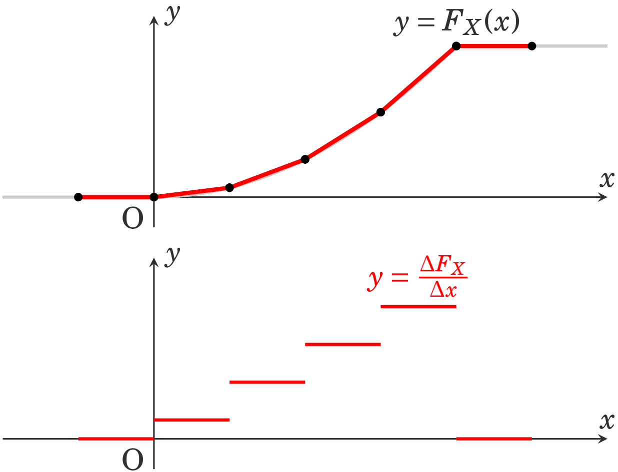
CDF to the density.



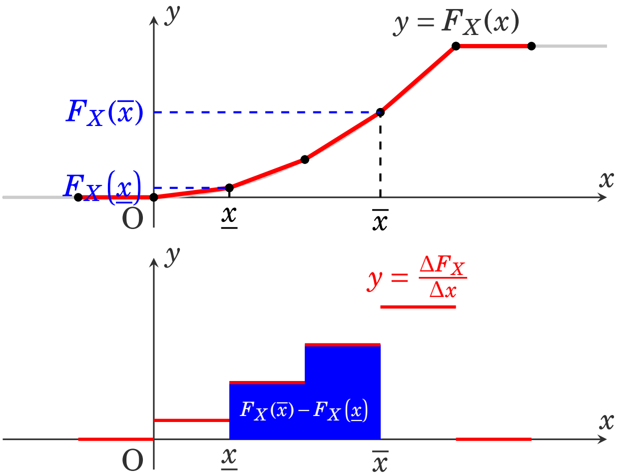
CDF to the density.



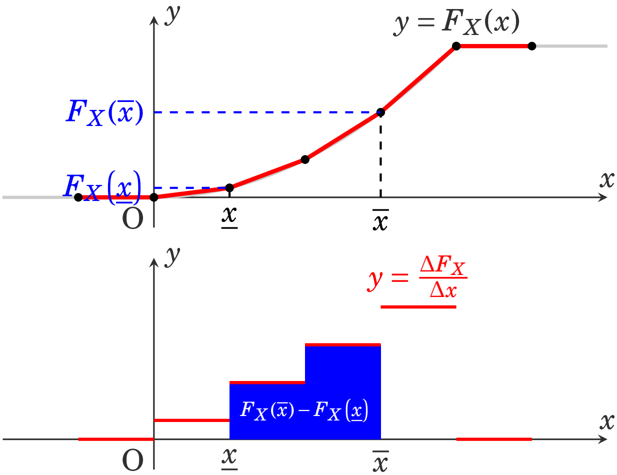
CDF to the density.



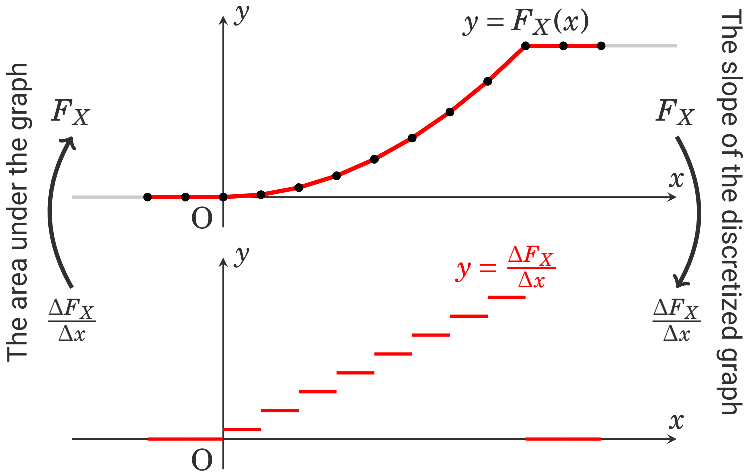
CDF to the density.



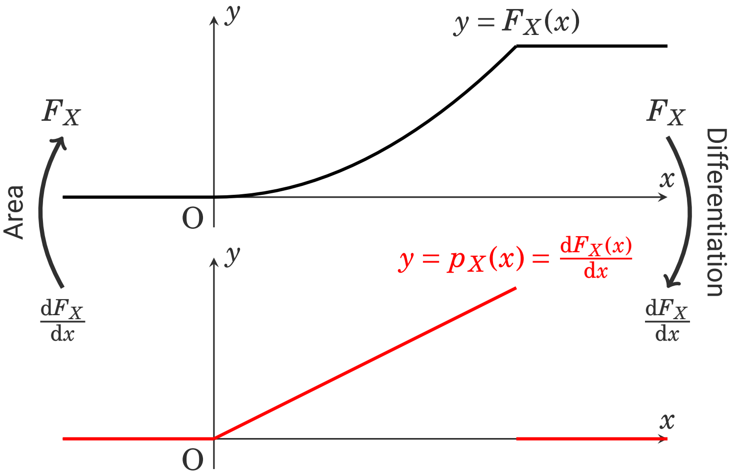
CDF to the density.



CDF to the density.



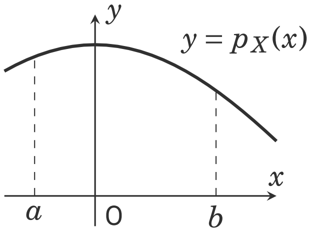
CDF to the density.



Probability density function (PDF)

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 $x = b$.

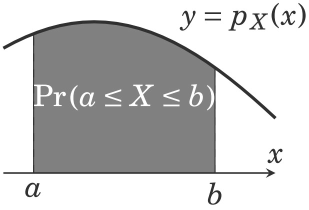


If the probability density function (PDF) of a random variable is given, the probability $\Pr(a \leq X \leq b)$ is given by the area under the PDF in the domain $[a, b]$.

Probability density function (PDF)

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 $[a, b]$.

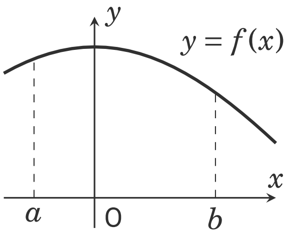
Outline

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 $\displaystyle\int_a^b f(x)\,\mathrm{d}x$.

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



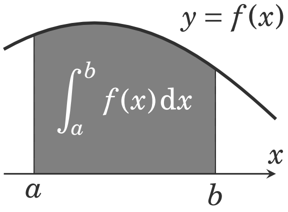
The definite integral is the area bounded by the graph of the function.

Definite Integral

Suppose that $a \le b$.

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

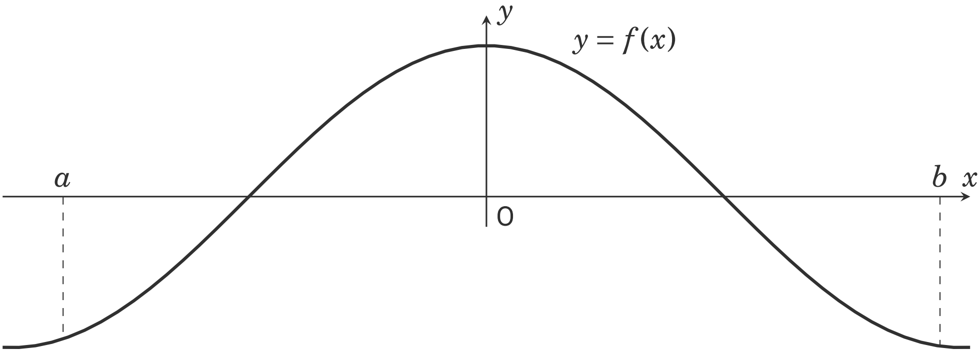
We also define $\int_b^a f(x)\,\mathrm{d}x := -\int_a^b f(x)\,\mathrm{d}x$.



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

Definite Integral: When the function takes negative values

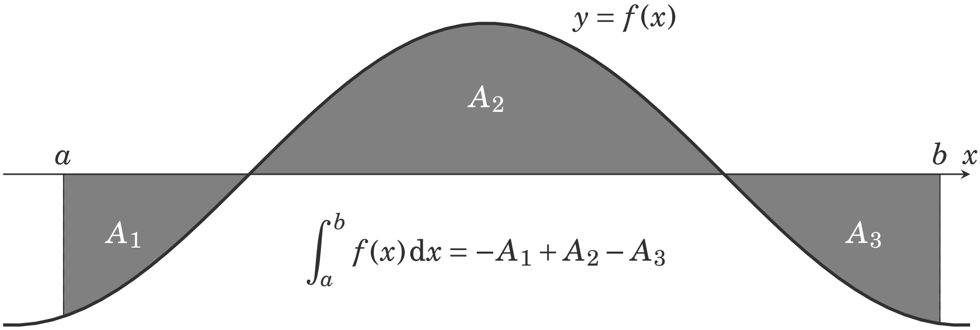
Areas bounded by the graph taking negative values are counted as negative values.



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.

- $\displaystyle\int_a^{+\infty} f(x)\,\mathrm{d}x := \lim_{b\to+\infty}\int_a^b f(x)\,\mathrm{d}x.$
- $\displaystyle\int_{-\infty}^{b} f(x)\,\mathrm{d}x := \lim_{a\to-\infty}\int_a^b f(x)\,\mathrm{d}x.$
- $\displaystyle\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x := \int_{-\infty}^{c} f(x)\,\mathrm{d}x + \int_c^{+\infty} f(x)\,\mathrm{d}x$, where c is an arbitrary real value[6].

[6] 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.

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

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

$$x(t) = \int_0^t v(t)\,\mathrm{d}t\,. \tag{37}$$

Outline

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

$$\int_{-\infty}^{+\infty} x p(x) \, \mathrm{d}x \, . \tag{38}$$

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

$$\sum_{x \in \mathcal{X}} x P(x), \tag{39}$$

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

$$\int_{-\infty}^{+\infty} f(x)p(x)\,\mathrm{d}x\,. \tag{40}$$

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

$$\sum_{x \in \mathcal{X}} f(x)P(x), \tag{41}$$

where P is the probability mass function.

The expectation does not always exist

If the PDF of a random variable X is given by

$$p(x) = \frac{1}{1+x^2}, \tag{42}$$

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

$$\int_{-\infty}^{+\infty} x p(x) \, \mathrm{d}x := \lim_{a \to -\infty} \int_a^c x p(x) \, \mathrm{d}x + \lim_{b \to +\infty} \int_c^b x p(x) \, \mathrm{d}x \tag{43}$$

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

$$\int_{-\infty}^{+\infty} (x-m)^2 p(x)\,\mathrm{d}x\,. \tag{44}$$

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

Outline

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

$$\int_a^x f(t)\,\mathrm{d}t = F(x) - F(a). \tag{45}$$

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

Calculating the definite integral

To calculate the definite integral

$$\int_a^b f(x)\,\mathrm{d}x\,, \tag{46}$$

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.

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 $f(x) = x$.

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

A primitive function is not unique. **However**, 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**.

We denote those primitive functions by $\int f(x)\,\mathrm{d}x$. If F is a primitive function, then we write like

$$\int f(x)\,\mathrm{d}x = F(x) + C, \tag{47}$$

since the function F_0 defined by $F_0(x) := F(x) + C$ is again a primitive function for any constant C. Here, the constant C is called the ***constant of integration***.

Example

The function $F(x) = \frac{1}{2}x^2$ is a primitive function of $f(x) = x$ since $F'(x) = f(x)$. Hence, $\displaystyle\int f(x)\,\mathrm{d}x = \frac{1}{2}x^2 + C$.

Finding the primitive function is not always easy.

To calculate the derivative, we had many useful formulae. Let $f, g : \mathbb{R} \to \mathbb{R}$ be differentiable functions and $c \in \mathbb{R}$ be a constant, then, e.g.,

- $(f+g)' = f' + g'$ for the sum,
- $(cf)' = cf'$ for the scalar product,

and

- $(fg)' = f'g + fg'$ for the product,
- $(g \circ f)' = \left(g' \circ f\right) f'$ for the composition.

Recall that the composition $g \circ f$ is defined by $(g \circ f)(x) = g(f(x))$.

We have similar formulae for the sum and scalar product, but no similar formulae for the general product and the composition.

Integration formulae for the sum and product

If F and G are the primitive functions of f and g, respectively, the following formulae hold.

- $F+G$ is a primitive function of $f+g$,
- cF is a primitive of cf.

Example

In the following, C is the constant of integration.

- $\displaystyle\int (\cos x + x^2)\,\mathrm{d}x = \sin x + \frac{1}{3}x^3 + C.$
- $\displaystyle\int 5\exp(x)\,\mathrm{d}x = 5\exp(x) + C.$

However, there is no general formula to write a primitive function of the product fg or the composition $g \circ f$ by F and G.

Integration by parts

Let f and g be real functions and F and G be those primitive functions. While we cannot generally write the primitive function of the product fg only by F and G, the technique, called ***integration by parts***, based on the following equation might help.

$$\int f(x)g(x)\,\mathrm{d}x = f(x)G(x) - \int f'(x)G(x)\,\mathrm{d}x\,. \tag{48}$$

Note that we assume that f is differentiable in the above.

By the above equation, we can find the primitive function of fg as long as we know that of $f'G$.

The proof of the above equation is easy if we differentiate the RHS.

Integration by parts

Let f and g be real functions and F and G be those primitive functions. While we cannot generally write the primitive function of the product fg only by F and G, the technique, called ***integration by parts***, based on the following equation might help.

$$\int f(x)g(x)\,\mathrm{d}x = f(x)G(x) - \int f'(x)G(x)\,\mathrm{d}x\,. \tag{48}$$

Example

$$\begin{aligned}\int x\cos(x)\,\mathrm{d}x &= x\sin(x) - \int (x)'\sin(x)\,\mathrm{d}x \\ &= x\sin(x) - \int 1\cdot\sin(x)\,\mathrm{d}x \\ &= x\sin(x) - (-\cos x) + C.\end{aligned} \tag{49}$$

Integration by parts

Let f and g be real functions and F and G be those primitive functions. While we cannot generally write the primitive function of the product fg only by F and G, the technique, called ***integration by parts***, based on the following equation might help.

$$\int f(x)g(x)\,\mathrm{d}x = f(x)G(x) - \int f'(x)G(x)\,\mathrm{d}x\,. \tag{48}$$

Example

$$\begin{aligned}\int \log(x)\,\mathrm{d}x = \int \log(x)\cdot 1\,\mathrm{d}x &= \log(x)\cdot x - \int (\log(x))'\cdot x\,\mathrm{d}x \\ &= \log(x)\cdot x - \int \frac{1}{x}\cdot x\,\mathrm{d}x \\ &= x\log(x) - x + C\,.\end{aligned} \tag{49}$$

Integration by substitution

Let f and g be real functions and assume f be differentiable. If the integrand includes the composition $g \circ f$, we cannot generally write the primitive function only by the primitive functions of f and g. However, we may find it by the following technique, called ***integration by substitution***.

Theorem (Integration by substitution for indefinite integral)

$$\int g(f(t))f'(t)\,\mathrm{d}t = \int g(x)\,\mathrm{d}x\bigg|_{x=f(t)}, \tag{50}$$

where the RHS means the function we obtain by substituting $x = f(t)$ to a primitive function of g.

Both directions of the above equation are useful.

Integration by substitution

Let f and g be real functions and assume f be differentiable. If the integrand includes the composition $g \circ f$, we cannot generally write the primitive function only by the primitive functions of f and g. However, we may find it by the following technique, called ***integration by substitution***.

Theorem (Integration by substitution for definite integral)

$$\int_a^b g(f(t))f'(t)\,\mathrm{d}t = \int_{f(a)}^{f(b)} g(x)\,\mathrm{d}x\,, \tag{50}$$

where the RHS means the function we obtain by substituting $x = f(t)$ for a primitive function of g.

Both directions of the above equation are useful.

Why do we call it integration by substitution?

The previous page's formula is called integration by substitution because the formula is informally given by substituting $x = f(t)$ as follows.

$$\begin{aligned}\int_a^b g(f(t))f'(t)\,\mathrm{d}t &= \int_{t=a}^{t=b} g(f(t))\frac{\mathrm{d}f(t)}{\mathrm{d}t}\,\mathrm{d}t \\ &= \int_{t=a}^{t=b} g(x)\frac{\mathrm{d}x}{\mathrm{d}t}\,\mathrm{d}t \\ &= \int_{t=a}^{t=b} g(x)\,\mathrm{d}x \\ &= \int_{x=f(a)}^{x=f(b)} g(x)\,\mathrm{d}x\,.\end{aligned} \tag{51}$$

Note that the above discussion is mathematically inaccurate (especially where we used $\frac{\mathrm{d}x}{\mathrm{d}t}\,\mathrm{d}t = \mathrm{d}x$). If we want to formally prove the formula, we should simply differentiate both sides of the formula for indefinite integral.

Examples of integration by substitution.

Recall the formula.

$$\int_a^b g(f(t))f'(t)\,\mathrm{d}t = \int_{f(a)}^{f(b)} g(x)\,\mathrm{d}x\,, \tag{52}$$

Example (integration by substitution: from left to right)

$$\begin{aligned}
\int_0^{+2} t\exp\left(-t^2\right)\mathrm{d}t &= -\frac{1}{2}\int_0^{+2} \exp\left(-t^2\right)\cdot(-2t)\,\mathrm{d}t \\
&= -\frac{1}{2}\int_0^{+2} \exp\left(-t^2\right)\cdot\left(-t^2\right)'\,\mathrm{d}t \\
&= -\frac{1}{2}\int_{-0^2}^{-2^2} \exp(x)\,\mathrm{d}x \\
&= -\frac{1}{2}[\exp(x)]_{-0^2}^{-2^2} = -\frac{1}{2}[\exp(-4)-\exp(0)] = \frac{1}{2}[1-\exp(-4)].
\end{aligned} \tag{53}$$

Examples of integration by substitution.

Recall the formula.

$$\int_a^b g(f(t))f'(t)\,\mathrm{d}t = \int_{f(a)}^{f(b)} g(x)\,\mathrm{d}x\,, \tag{52}$$

Example (integration by substitution: from right to left)

$$\begin{aligned}\int_0^1 \sqrt{1-x^2}\,\mathrm{d}x &= \int_\pi^0 \sqrt{1-\cos^2(t)}(\cos(t))'\,\mathrm{d}t \quad \text{since } \cos(\pi)=0, \cos(0)=1,\\ &= \int_\pi^0 \sqrt{1-\cos^2(t)}(-\sin(t))\,\mathrm{d}t \\ &= \int_0^\pi \sin^2(t)\,\mathrm{d}t = \int_0^\pi \frac{1-\cos(2t)}{2}\,\mathrm{d}t = \left[\frac{1}{2}t - \frac{1}{4}\sin(2t)\right]_0^\pi = \frac{1}{2}\pi.\end{aligned} \tag{53}$$

Finding a primitive function of the product and composition is not easy.

We know that the primitive functions of $\frac{1}{x}$ and $\sin$, or $\exp$ and $-x^2$. Indeed,

$$\int \frac{1}{x} \mathrm{d}x = \log|x| + C, \int \sin x \, \mathrm{d}x = -\cos + C, \int \left(-x^2\right) \mathrm{d}x = -\frac{1}{3}x^3 + C, \int \exp(x) \, \mathrm{d}x = \exp(x) + C \tag{54}$$

However, it is known that the primitive functions of $\frac{1}{x} \sin x$ and $\exp\left(-x^2\right)$ are not ***elementary***, although $\frac{1}{x} \sin x$ and $\exp\left(-x^2\right)$ themselves are elementary.

Here, we call a function ***elementary*** if we can write the function as a composition of finitely many

- algebraic functions, functions represented as a root of polynomial-function-coefficient polynomial equations, including polynomial, rational functions and fractional powers, e.g., $5x^2 + x - 3$, $\sqrt{3}x + 5$, $\frac{3x+1}{-2x^2+x+5}$, etc.
- trigonometric functions, e.g., $\sin x$, $\cos x$ etc.,
- exponential function $\exp x$,
- logarithmic function $\log x$.

Finding a primitive function of the product and composition is not easy.

We know that the primitive functions of $\frac{1}{x}$ and $\sin$, or $\exp$ and $-x^2$. Indeed,

$$\int \frac{1}{x}\,\mathrm{d}x = \log|x| + C, \int \sin x\,\mathrm{d}x = -\cos + C, \int \left(-x^2\right)\mathrm{d}x = -\frac{1}{3}x^3 + C, \int \exp(x)\,\mathrm{d}x = \exp(x) + \tag{54}$$

However, it is known that the primitive functions of $\frac{1}{x}\sin x$ and $\exp\left(-x^2\right)$ are not ***elementary***, although $\frac{1}{x}\sin x$ and $\exp\left(-x^2\right)$ themselves are elementary.

Roughly speaking, most functions we can imagine without the inverse function and the primitive function are elementary.

The fact that the primitive functions of $\frac{1}{x}\sin x$ and $\exp\left(-x^2\right)$ are not elementary means we have no way to write those primitive functions.

From the computer science viewpoint, the above fact means that we cannot easily find the exact value of the integrals of those functions. Some non-elementary primitive functions might be implemented by some libraries if they are famous. If they are not

Finding a primitive function of the product and composition is not easy.

We know that the primitive functions of $\frac{1}{x}$ and $\sin$, or $\exp$ and $-x^2$. Indeed,

$$\int \frac{1}{x} \mathrm{d}x = \log|x| + C, \int \sin x \, \mathrm{d}x = -\cos + C, \int \left(-x^2\right) \mathrm{d}x = -\frac{1}{3}x^3 + C, \int \exp(x) \, \mathrm{d}x = \exp(x) + C \tag{54}$$

However, it is known that the primitive functions of $\frac{1}{x} \sin x$ and $\exp\left(-x^2\right)$ are not ***elementary***, although $\frac{1}{x} \sin x$ and $\exp\left(-x^2\right)$ themselves are elementary.

In fact, these functions are important in many areas.

- The PDF of the normal distribution is proportional to $\exp\left(-x^2\right)$. The normal distribution is the most important distribution in probability theory, owing to the central limit theorem.
- The sine cardinal function $\frac{\sin x}{x}$ appears in many application areas, including physics, probability theory, signal processing, optics, etc., because it is the Fourier transform of the rectangle function.

Outline

Handling a continuous multivariate random variable

Similar to the univariate random variable case, we can define the cumulative distribution function (CDF) of the distribution.

Definition (The CDF of bivariate RV)

Let X and Y be random variables. The ***cumulative distribution function (CDF)*** $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ of X and Y is defined by

$$F_{X,Y}(x,y) := \Pr(X \leq x \wedge Y \leq y), \tag{55}$$

where ∧ indicates the logical "and" statement.

Using the CDF, we can calculate the probability $\Pr(a_1 < X \leq b_1 \wedge a_2 < X \leq b_2)$ by

$$\Pr(a_1 < X \leq b_1 \wedge a_2 < X \leq b_2) = F_{X,Y}(b_1,b_2) - F_{X,Y}(a_1,b_2) - F_{X,Y}(b_1,a_2) + F_{X,Y}(b_1,b_2) \tag{56}$$

Multiple integral

Let $D = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_m, b_m]$ be a m-dimensional hyper-rectangle. 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

$$\{(x_1, x_2, \ldots, x_m, f(x_1, x_2, \ldots, x_m)) | (x_1, x_2, \ldots, x_m) \in D\} \tag{57}$$

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 ***multiple 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 multiple integral of f on D by

$$\int_A f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} := \int_D 1_A(\boldsymbol{x}) f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,. \tag{58}$$

Integration by substitution for a double integral

Theorem (Integration by substitution for a double integral)

$$\int_U f\left(\varphi_1(u_1,u_2),\varphi_2(u_1,u_2)\right)\left|\det\left(\begin{bmatrix}\frac{\partial\varphi_1}{\partial u_1}(u_1,u_2) & \frac{\partial\varphi_1}{\partial u_2}(u_1,u_2)\\ \frac{\partial\varphi_2}{\partial u_1}(u_1,u_2) & \frac{\partial\varphi_2}{\partial u_2}(u_1,u_2)\end{bmatrix}\right)\right|\mathrm{d}u_1\,\mathrm{d}u_2$$
$$= \int_{\boldsymbol{\varphi}(U)} f(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \tag{59}$$

Here, recall that

$$\det\left(\begin{bmatrix}a & b\\ c & d\end{bmatrix}\right) = ad - bc. \tag{60}$$

An example of integration by substitution

Most practical substitutions are given by the polar coordinate: $x = r\cos\theta, y = r\sin\theta$.

By this substitution, we have that $\sqrt{x^2 + y} = r$.

Also, the determinant of the Jacobian of the coordinate transform is given by

$$\det\left(\begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix}\right) = \det\left(\begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}\right) = r\cos^2\theta - \left(-r\sin^2\right) = r. \tag{61}$$

Using the above results, we can calculate, for example,

$$\begin{aligned} \iint_{x^2+y^2\leq 1} \left(1 - \sqrt{x^2 + y^2}\right) \mathrm{d}x\,\mathrm{d}y &= \int_0^{2\pi}\int_0^1 (1-r)\left|\det\left(\begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix}\right)\right| \mathrm{d}r\,\mathrm{d}\theta \\ &= \int_0^{2\pi}\int_0^1 (1-r)|r|\,\mathrm{d}r\,\mathrm{d}\theta \\ &= \int_0^{2\pi}\left[\int_0^1 \left(r - r^2\right)\mathrm{d}r\right]\mathrm{d}\theta = \int_0^{2\pi} \frac{1}{6}\,\mathrm{d}\theta = \frac{1}{3}\pi. \end{aligned} \tag{62}$$

Integration by substitution for a multiple integral

Theorem (Integration by substitution for a multiple integral)

$$\int_U f\left(\boldsymbol{\varphi}(\boldsymbol{u})\right)\left|\det\left(\frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{u}}(\boldsymbol{u})\right)\right| \mathrm{d}\boldsymbol{u} = \int_{\boldsymbol{\varphi}(U)} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{63}$$

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

$$\Pr((X_1, X_2, \ldots, X_m) \in A) = \int_A p_{X_1,X_2,\ldots,X_m}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \tag{64}$$

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

Joint PDF examples

Example (Uniform distribution)

Let X and Y be RVs following the bivariate uniform distribution with the support $[0,3] \times [-1,+1]$. The RVs X and Y have has the joint PDF

$$p_{X,Y} = \begin{cases} \frac{1}{6} & \text{if } (x,y) \in [0,3] \times [-1,+1], \\ 0 & \text{if } (x,y) \notin [0,3] \times [-1,+1]. \end{cases} \tag{65}$$

The probability $\Pr\big((X,Y) \in [0,\frac{1}{2}] \times [0,\frac{1}{4}]\big)$ is given by

$$\int_0^{\frac{1}{4}} \int_0^{\frac{1}{2}} \frac{1}{6} \,\mathrm{d}x \,\mathrm{d}y = \int_0^{\frac{1}{4}} \left[\frac{1}{6}x\right]_0^{\frac{1}{2}} \mathrm{d}y = \int_0^{\frac{1}{4}} \frac{1}{12} \,\mathrm{d}y = \left[\frac{1}{12}y\right]_0^{\frac{1}{4}} = \frac{1}{48} \tag{66}$$

Joint PDF examples

Example (Bivariate normal distribution)

Let X and Y be RVs following the bivariate uniform distribution with the support $[0,3] \times [-1,+1]$. The RVs X and Y have has the joint PDF

$$p_{X,Y} = \begin{cases} \frac{1}{6} & \text{if } (x,y) \in [0,3] \times [-1,+1], \\ 0 & \text{if } (x,y) \notin [0,3] \times [-1,+1]. \end{cases} \tag{65}$$

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

$$\begin{aligned} p_X(x) &= \int_{-\infty}^{+\infty} p_{X,Y}(x,y)\mathrm{d}y, \\ p_Y(y) &= \int_{-\infty}^{+\infty} p_{X,Y}(x,y)\mathrm{d}x. \end{aligned} \tag{66}$$

Conditional PDF (bivariate cases)

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

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}, \quad p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}. \tag{67}$$

Independency

Let X and Y be RVs and assume that they have a joint PDF $p_{X,Y}$ and let their marginal PDFs be p_X and p_Y. Also, denote the conditional PDF of X given Y and that of Y given X by $p_{X|Y}$ and $p_{Y|X}$, respectively.

We say that the RVs X and Y are (mutually) independent if one of the following equivalent conditions holds

- $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for all (x,y).
- $p_{X|Y}(x|y) = p_X(x)$ for all (x,y) such that $p_Y(y) \neq 0$.
- $p_{Y|X}(y|x) = p_Y(y)$ for all (x,y) such that $p_X(x) \neq 0$.

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

$$\int_{\mathbb{R}^m} f(\boldsymbol{x}) p_{X_1, X_2, \ldots, X_m}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{68}$$

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

$$\mathrm{Cov}(X,Y) = \int_{\mathbb{R}^2} (x - \mu_X)(y - \mu_Y) p_{X,Y}(x,y) \mathrm{d}x \mathrm{d}y \tag{69}$$

Calculating multi integral by iterated integral

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

Theorem

Under some loose conditions[7]*, we have that*

$$\begin{aligned}
&\iint_A p(x, y) \, \mathrm{d}x \, \mathrm{d}y \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} 1_A(x, y) p(x, y) \, \mathrm{d}x \right] \mathrm{d}y \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} 1_A(x, y) p(x, y) \, \mathrm{d}y \right] \mathrm{d}x \, .
\end{aligned} \tag{70}$$

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

Outline

Exercise (CDF of a continuous RV)

Let X be a random variable whose CDF is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x \le 0, \\ x & \text{if } 0 \le x \le 1, \\ 1 & \text{if } x \ge 1. \end{cases} \tag{71}$$

Find the probability of $0.2 \le X \le 0.7$.

Exercise (CDF to PDF)

Let X be a random variable whose CDF is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ \frac{1}{4}x^2 & \text{if } 0 \leq x \leq 2, \\ 1 & \text{if } x \geq 2. \end{cases} \tag{72}$$

Find the PDF p_X.

Exercise (PDF to probability)

Let X be a random variable whose PDF is given by

$$p_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ \exp(-x) & \text{if } x \geq 0. \end{cases} \tag{73}$$

Find the probability of the event $1 \leq X \leq 2$.

Exercise (Indefinite integral)

Find the following indefinite integrals

- $\int \left(\cos x + x^2\right) \mathrm{d}x$
- $\int 5 \exp(x) \, \mathrm{d}x$
- $\int x \cos(x) \, \mathrm{d}x$
- $\int \log(x) \, \mathrm{d}x$

Exercise (Definite integral)

Find the following definite integrals

- $\displaystyle\int_{\pi}^{2\pi} \left(\cos x + x^2\right) \mathrm{d}x$
- $\displaystyle\int_{0}^{1} 5\exp(x)\,\mathrm{d}x$
- $\displaystyle\int_{\pi}^{0} x\cos(x)\,\mathrm{d}x$
- $\displaystyle\int_{1}^{e} \log(x)\,\mathrm{d}x$
- $\displaystyle\int_{0}^{+2} t\exp\left(-t^2\right) \mathrm{d}t$
- $\displaystyle\int_{0}^{1} \sqrt{1-x^2}\,\mathrm{d}x$

Exercise (Improper integral)

Find the following improper integrals

- $\displaystyle\int_0^{+\infty} \exp(-x)\,\mathrm{d}x.$
- $\displaystyle\int_{-\infty}^{+\infty} \frac{1}{1+x^2}\,\mathrm{d}x.$

You can use the fact that $\frac{\mathrm{d}}{\mathrm{d}x}\arctan(x) = \frac{1}{1+x^2}$, where $\arctan$ is the inverse function of $\tan$ with the domain limited to $(-\pi, +\pi)$.

Exercise (Double integral)

- Evaluate the definite integral $\iint_D 3x^2 y \, \mathrm{d}x \, \mathrm{d}y$ where D is defined by $0 \leq x \leq 1$ and $0 \leq y \leq 2$.
- Evaluate the definite integral $\iint_D \left(x^2 + y^2\right) \mathrm{d}x \, \mathrm{d}y$ where D is the region bounded by the curves $y = x$, $y = 2x$, $x = 1$, and $x = 2$.

Exercise (Integration by substitution for a double integral)

Evaluate the following values

- $\displaystyle\iint_{x^2+y^2\leq 1}\left(1-\sqrt{x^2+y^2}\right)\mathrm{d}x\,\mathrm{d}y$,
- $\displaystyle\iint_{x^2+y^2\leq 1}\exp\left(-\sqrt{x^2+y^2}\right)\mathrm{d}x\,\mathrm{d}y$,
- $\displaystyle\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\exp\left(-\left(x^2+y^2\right)\right)\mathrm{d}x\,\mathrm{d}y$.

Outline

4. Sample Statistics

    4.1 Introduction: why do we learn sample statistics?

    4.2 Terminology

    4.3 Sample mean, law of large numbers, and central limit theorem

    4.4 Estimation of distribution and parametric model

    4.5 Likelihood

    4.6 Maximum likelihood estimator

    4.7 Exercises

Outline

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**.

Learning outcomes

By the end of this topic, you should be able to:

- Explain the difference between summary statistics and sample statistics,
- Estimate the true mean of an unknown distribution by finite size sample,
- Explain why many random variables in the real world follow a normal distribution, and
- Estimate an unknown distribution using a parametric model and maximum likelihood estimator.

Outline

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 in real applications, so sample statistics are practically necessary to handle probability.

Outline

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***

$$m_m = \frac{1}{m}(X_1 + X_2 + \cdots + X_m), \tag{74}$$

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 $X_1, X_2, \ldots$ 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

$$\overline{X}_m := \frac{1}{m}(X_1 + X_2 + \cdots + X_m). \tag{75}$$

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

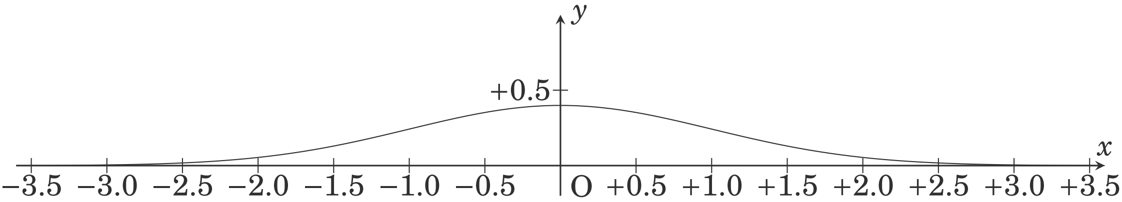
- The difference between the sample mean and the true expectation is proportional to the standard deviation σ of the true distribution and $\frac{1}{\sqrt{m}}$,
- With appropriate scaling, the distribution of the sample mean converges to a ***normal distribution (Gaussian distribution)***,

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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{76}$$



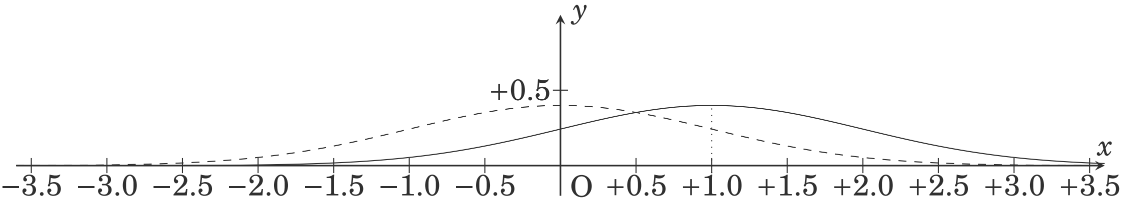
Normal distributions' PDF ($\mu = 0, \sigma = 1$).

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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right). \tag{76}$$



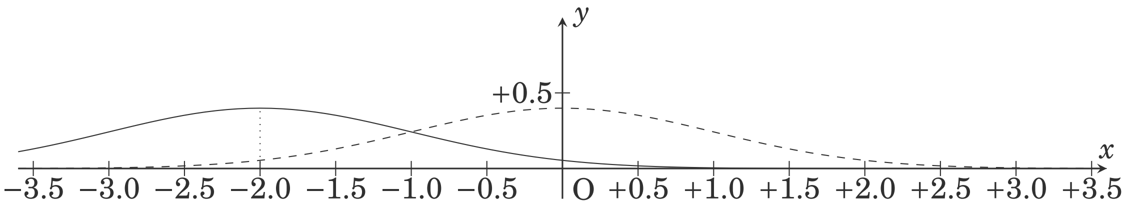
Normal distributions' PDF (Solid: $\mu = 1, \sigma = 1$, Dashed: $\mu = 0, \sigma = 1$).

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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{76}$$



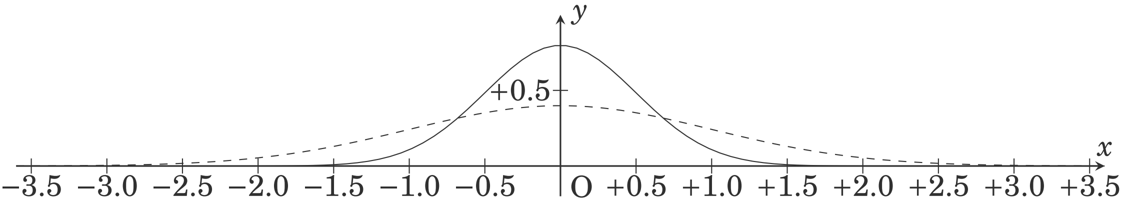
Normal distributions' PDF (Solid: $\mu = -2, \sigma = 1$, Dashed: $\mu = 0, \sigma = 1$).

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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right). \tag{76}$$



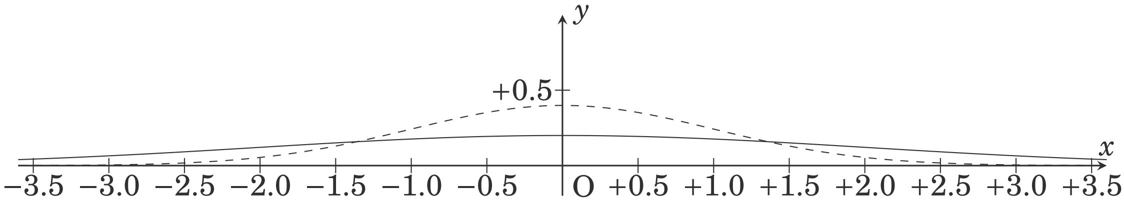
Normal distributions' PDF (Solid: $\mu = 0, \sigma = 0.5$, Dashed: $\mu = 0, \sigma = 1$).

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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{76}$$



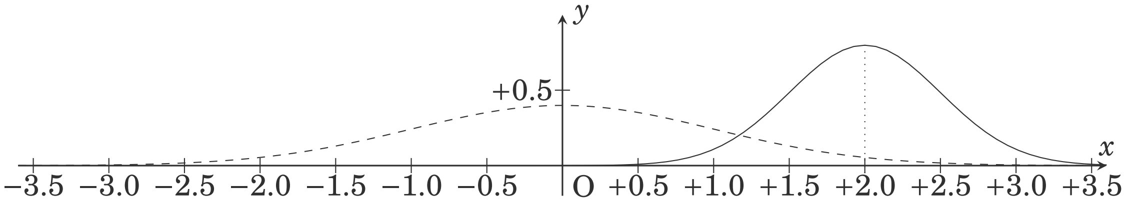
Normal distributions' PDF (Solid: $\mu = 0, \sigma = 2.0$, Dashed: $\mu = 0, \sigma = 1$).

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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{76}$$



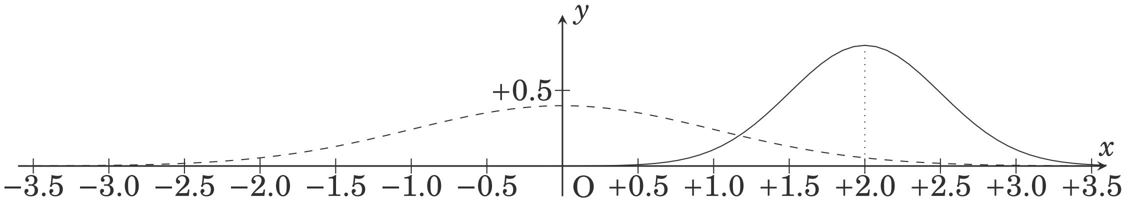
Normal distributions' PDF (Solid: $\mu = 2, \sigma = 0.5$, Dashed: $\mu = 0, \sigma = 1$).

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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{76}$$



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

$$\overline{X}_m := \frac{1}{m}(X_1 + X_2 + \cdots + X_m). \tag{77}$$

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**.

Example of the convergence by the CLT

Example

Let $X_1, X_2, \ldots$ be an infinite sequence of independently identically distributed RVs, where X_i takes $+1$ or -1 with probability $\frac{1}{2}$ for each.
Then the mean and the variance of X_i are 0 and 1, respectively.
According to the CLT, the CDF of $\sqrt{m}\overline{X_m}$ converges to that of the standard normal distribution $\mathcal{N}(0, 1)$.
Note that $\overline{X_m}$ is a discrete random variable since each X_i is. Hence, the random variable $\sqrt{m}\overline{X_m}$ does not have a PDF. Therefore, we **CANNOT** say that the PDF of $\sqrt{m}\overline{X_m}$ converges to that of the standard normal distribution.

Outline

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 an infinite set[8], 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.

[8] This is almost always the case if we consider a continuous RV

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.

Parametric model examples

Example (Bernoulli distribution)

The Bernoulli distribution[9] is a discrete parametric model, whose parameter is usually denoted by θ. The support and the parameter set are $\mathscr{X} = \{0, 1\}$ and $\Theta = [0, 1]$, respectively. The parametrized PMF $P(x;\theta)$ is given by $P(1;\theta) = \theta$. Thus, we have $P(0;\theta) = 1 - \theta$.

[9] A parametric model is often called like the XXX distribution, but it is, indeed, a parametrized **set** of distributions.

Parametric model examples

Example (Normal distribution)

The normal distribution is a continuous parametric model, which has the mean parameter $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_{>0}$. That is, the parameter set is $\Theta = \mathbb{R} \times \mathbb{R}_{>0}$. The parametrized PDF $p\left(x; \mu, \sigma^2\right)$ is given by $p\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x - \mu\right)\right)$.

Outline

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.

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{78}$$

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 $\mathit{\Theta}$ 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.

$$p(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot p(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot p(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{78}$$

Examples of likelihood calculation

Suppose that we have data points $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, and consider the Bernoulli distribution $P(0;\theta) = 1-\theta, P(1;\theta) = \theta$.

The likelihood of the Bernoulli distribution with θ on the data is given by

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta\cdot\theta\cdot(1-\theta)\cdot\theta. \quad (79)$$

- The likelihood of $\theta = 0$ is $0 \cdot 0 \cdot (1-0) \cdot 0 = 0$.

Examples of likelihood calculation

Suppose that we have data points $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, and consider the Bernoulli distribution $P(0;\theta) = 1-\theta, P(1;\theta) = \theta$.

The likelihood of the Bernoulli distribution with θ on the data is given by

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta\cdot\theta\cdot(1-\theta)\cdot\theta. \quad (79)$$

- The likelihood of $\theta = 0$ is $0 \cdot 0 \cdot (1-0) \cdot 0 = 0$.
- The likelihood of $\theta = \frac{1}{4}$ is $\frac{1}{4} \cdot \frac{1}{4} \cdot \left(1 - \frac{1}{4}\right) \cdot \frac{1}{4} = \frac{3}{256}$.

Examples of likelihood calculation

Suppose that we have data points $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, and consider the Bernoulli distribution $P(0;\theta) = 1-\theta, P(1;\theta) = \theta$.

The likelihood of the Bernoulli distribution with θ on the data is given by

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot \theta. \quad (79)$$

- The likelihood of $\theta = 0$ is $0 \cdot 0 \cdot (1-0) \cdot 0 = 0$.
- The likelihood of $\theta = \frac{1}{4}$ is $\frac{1}{4} \cdot \frac{1}{4} \cdot \left(1 - \frac{1}{4}\right) \cdot \frac{1}{4} = \frac{3}{256}$.
- The likelihood of $\theta = \frac{1}{2}$ is $\frac{1}{2} \cdot \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) \cdot \frac{1}{2} = \frac{1}{16} = \frac{16}{256}$.

Examples of likelihood calculation

Suppose that we have data points $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, and consider the Bernoulli distribution $P(0;\theta) = 1 - \theta, P(1;\theta) = \theta$.

The likelihood of the Bernoulli distribution with θ on the data is given by

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta\cdot\theta\cdot(1-\theta)\cdot\theta. \quad (79)$$

- The likelihood of $\theta = 0$ is $0 \cdot 0 \cdot (1 - 0) \cdot 0 = 0$.
- The likelihood of $\theta = \frac{1}{4}$ is $\frac{1}{4} \cdot \frac{1}{4} \cdot \left(1 - \frac{1}{4}\right) \cdot \frac{1}{4} = \frac{3}{256}$.
- The likelihood of $\theta = \frac{1}{2}$ is $\frac{1}{2} \cdot \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) \cdot \frac{1}{2} = \frac{1}{16} = \frac{16}{256}$.
- The likelihood of $\theta = \frac{3}{4}$ is $\frac{3}{4} \cdot \frac{3}{4} \cdot \left(1 - \frac{3}{4}\right) \cdot \frac{3}{4} = \frac{27}{256}$.

Examples of likelihood calculation

Suppose that we have data points $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, and consider the Bernoulli distribution $P(0;\theta) = 1-\theta, P(1;\theta) = \theta$.

The likelihood of the Bernoulli distribution with θ on the data is given by

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta\cdot\theta\cdot(1-\theta)\cdot\theta. \quad (79)$$

- The likelihood of $\theta = 0$ is $0 \cdot 0 \cdot (1-0) \cdot 0 = 0$.
- The likelihood of $\theta = \frac{1}{4}$ is $\frac{1}{4} \cdot \frac{1}{4} \cdot \left(1 - \frac{1}{4}\right) \cdot \frac{1}{4} = \frac{3}{256}$.
- The likelihood of $\theta = \frac{1}{2}$ is $\frac{1}{2} \cdot \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) \cdot \frac{1}{2} = \frac{1}{16} = \frac{16}{256}$.
- The likelihood of $\theta = \frac{3}{4}$ is $\frac{3}{4} \cdot \frac{3}{4} \cdot \left(1 - \frac{3}{4}\right) \cdot \frac{3}{4} = \frac{27}{256}$.
- The likelihood of $\theta = 1$ is $1 \cdot 1 \cdot (1-1) \cdot 1 = 0$.

Examples of likelihood calculation

Suppose that we have data points $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, and consider the Bernoulli distribution $P(0;\theta) = 1 - \theta, P(1;\theta) = \theta$.

The likelihood of the Bernoulli distribution with θ on the data is given by

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta \cdot \theta \cdot (1 - \theta) \cdot \theta. \quad (79)$$

- The likelihood of $\theta = 0$ is $0 \cdot 0 \cdot (1 - 0) \cdot 0 = 0$.
- The likelihood of $\theta = \frac{1}{4}$ is $\frac{1}{4} \cdot \frac{1}{4} \cdot \left(1 - \frac{1}{4}\right) \cdot \frac{1}{4} = \frac{3}{256}$.
- The likelihood of $\theta = \frac{1}{2}$ is $\frac{1}{2} \cdot \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) \cdot \frac{1}{2} = \frac{1}{16} = \frac{16}{256}$.
- The likelihood of $\theta = \frac{3}{4}$ is $\frac{3}{4} \cdot \frac{3}{4} \cdot \left(1 - \frac{3}{4}\right) \cdot \frac{3}{4} = \frac{27}{256}$.
- The likelihood of $\theta = 1$ is $1 \cdot 1 \cdot (1 - 1) \cdot 1 = 0$.

Hence, among the above three, the distribution given by $\theta = \frac{3}{4}$ most likely generates the data sequence $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$.

Probability and likelihood

The value of the product

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}) \tag{80}$$

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).

Outline

Maximum likelihood estimator

Once we define the likelihood of a distribution, all we need to do is 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

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{81}$$

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[9].

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

$$P(\boldsymbol{x}_1;\boldsymbol{\theta})\cdot P(\boldsymbol{x}_2;\boldsymbol{\theta})\cdot\cdots\cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{82}$$

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

$$\log P(\boldsymbol{x}_1;\boldsymbol{\theta})+\log P(\boldsymbol{x}_2;\boldsymbol{\theta})+\cdots+\log P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{83}$$

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

$$-\log P(\boldsymbol{x}_1;\boldsymbol{\theta})-\log P(\boldsymbol{x}_2;\boldsymbol{\theta})-\cdots-\log P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{84}$$

---

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

Why do we consider the logarithm of the likelihood?

- The likelihood is a product and its logarithm is a sum. When we maximize it in a computer, we rely on its derivative (gradient descent methods). Differentiation of a sum is much easier than that of a product, so the (negative) log-likelihood has an advantage over the original likelihood from the optimization viewpoint.
- If the data size m is large, the absolute value of the likelihood, the product of many small values, tends to be too small to represent in a computer (underflow). Since the logarithm sees the power index, it can handle extremely small likelihood.
- The negative log-likelihood can be interpreted as the sum of the errors. For example, we can interpret the negative log-likelihood of the normal distribution as the squared error.

The MLE of the normal distribution minimizes the square 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

$$\begin{aligned}
&\log\left(2\pi\sigma^2\right) + \frac{\left(x_1-\mu\right)^2}{2\sigma^2} + \log\left(2\pi\sigma^2\right) + \frac{\left(x_2-\mu\right)^2}{2\sigma^2} + \cdots + \log\left(2\pi\sigma^2\right) + \frac{\left(x_m-\mu\right)^2}{2\sigma^2} \\
&= m\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left[\left(x_1-\mu\right)^2 + \left(x_2-\mu\right)^2 + \cdots + \left(x_m-\mu\right)^2\right].
\end{aligned} \tag{85}$$

The MLE of the normal distribution minimizes the square 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

$$\begin{aligned} &\log\left(2\pi\sigma^2\right) + \frac{\left(x_1-\mu\right)^2}{2\sigma^2} + \log\left(2\pi\sigma^2\right) + \frac{\left(x_2-\mu\right)^2}{2\sigma^2} + \cdots + \log\left(2\pi\sigma^2\right) + \frac{\left(x_m-\mu\right)^2}{2\sigma^2} \\ &= m\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left[\left(x_1-\mu\right)^2 + \left(x_2-\mu\right)^2 + \cdots + \left(x_m-\mu\right)^2\right]. \end{aligned} \tag{85}$$

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.

MLE example: Bernoulli case

Example

Suppose that we have data points $x_1, x_2, \ldots, x_m$, and consider the Bernoulli distribution $P(0;\theta) = 1-\theta, P(1;\theta) = \theta$.
The negative log-likelihood of the Bernoulli distribution with θ on the data is given by

$$-\log P(x_1;\theta)P(x_2;\theta)\ldots P(x_m;\theta) = m_0 \log(1-\theta) + m_1 \log\theta, \tag{86}$$

where m_0 and m_1 are the numbers of zeros and ones in the data sequence. Obviously, $m_0 + m_1 = m$, and the sample mean $\overline{x} = \frac{m_1}{m}$. Let l denote the above negative log-likelihood.
Suppose that $m_0 \neq 0$ and $m_1 \neq 0$, then l takes the minimum[10] if and only if $\theta = \frac{m_1}{m} = \overline{x}$.
Hence, the MLE $\hat{\theta} = \frac{m_1}{m} = \overline{x}$.
For example, if $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, then $\hat{\theta} = \frac{m_1}{m} = \overline{x} = \frac{3}{4}$.

---

[10] To prove it, differentiate the loss by θ and apply the first derivative test.

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 $\frac{1}{\sqrt{m}}$.

Outline

Exercise (Standard normal distribution)

Write down the standard normal distribution’s (probability density function) PDF.

Exercise (The central limit theorem (CLT))

Let $X_1, X_2, \ldots$ be an infinite sequence of independently identically distributed RVs. Assume the distribution of each random variable X_i is given by one of the following. For each case, apply the CLT and find what random variable converges to which normal distribution.

- The probability mass function P_{X_i} defined by $P_{X_i}(-1) = \frac{1}{4}$ and $P_{X_i}(+1) = \frac{3}{4}$.
- The probability density function p_{X_i} defined by $p_{X_i}(x) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-3}{2}\right)^2\right)$.

Exercise (Exercise: likelihood calculation)

Suppose that we have data points $(x_1, x_2, x_3, x_4) = (1, 0, 0, 1)$, and consider the Bernoulli distribution $P(0;\theta) = 1 - \theta, P(1;\theta) = \theta$.

Find the likelihoods of the Bernoulli distribution given by $\theta = 0, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1$.

Also, answer which distribution most likely generates the data points.

Outline

5. Statistical Test
   5.1 Introduction: why do we learn statistical tests?
   5.2 The logic of statistical tests
   5.3 Example test statistics
   5.4 p-value
   5.5 Failure of statistical test
   5.6 Exercises

Outline

Statistical tests support our judgements

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 $\theta \in \mathscr{H}_1$ or not.

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

Outline

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 $\theta \in \mathscr{H}_1$,” is:

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

The hypothesis $\mathscr{H}_0 \coloneqq \mathscr{H} \setminus \mathscr{H}_1$ is called the null hypothesis. Here, $\mathscr{H}$ 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***.

Outline

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***.

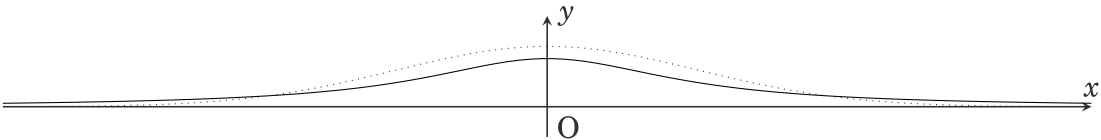
$$t := \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{m}}}, \tag{87}$$

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

$$\overline{X} := \frac{1}{m} \sum_{i=1}^{m} X_i, \quad s := \sqrt{\frac{1}{m} \left(X_i - \overline{X} \right)^2}. \tag{88}$$

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.



Black solid curve: the PDF of the t-distribution with 1 degree of freedom.

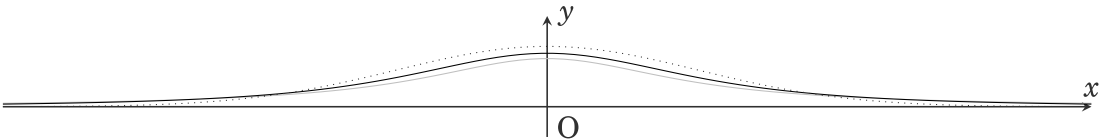
Black dotted curve: the PDF of the standard normal distribution.

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.



Black solid curve: the PDF of the t-distribution with 2 degree of freedom.

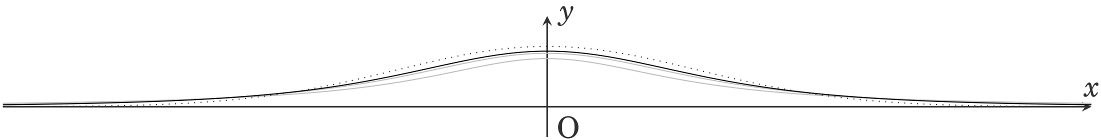
Black dotted curve: the PDF of the standard normal distribution.

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.



Black solid curve: the PDF of the t-distribution with 3 degree of freedom.

Black dotted curve: the PDF of the standard normal distribution.

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

$$p_{m-1}(x) = \frac{\Gamma\left(\frac{m}{2}\right)}{\sqrt{(m-1)\pi}\,\Gamma\left(\frac{m-1}{2}\right)}\left(1+\frac{x^2}{m-1}\right)^{-\frac{m}{2}} \tag{89}$$

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

Outline

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 [11]. The probability is called the ***p-value***. A small p-value indicates that the value of the statistic takes an extreme value.

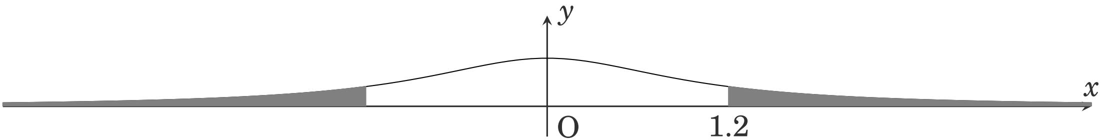
[11] Hence, we need to define in which case the value of the statistic is extreme. Although it is intuitive for well-known cases, there does not seem to be a way to mathematically decide it.

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

$$p = \Pr(|t| > |t_0|). \tag{90}$$



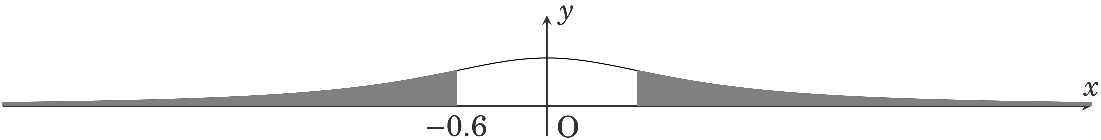
The p-value (the gray area) when t takes $t_0 = 1.2$.

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

$$p = \Pr(|t| > |t_0|). \tag{90}$$



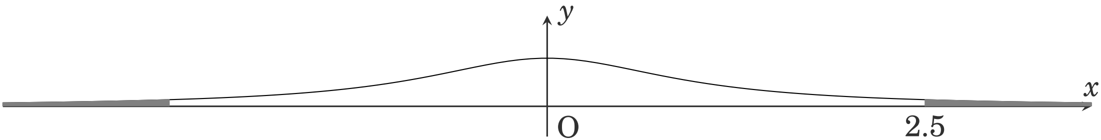
The p-value (the gray area) when t takes $t_0 = -0.6$.

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

$$p = \Pr(|t| > |t_0|). \tag{90}$$



The p-value (the gray area) when t takes $t_0 = 2.5$.

Significance level

We reject a hypothesis consisting of a single distribution if the p-value of the distribution on the data points is small[12].

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 $\mathscr{H}_1$.

---

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

Statistical test procedure

The standard procedure of the statistical test is the following.

- **Step 1**:
- **Step 2**:
- **Step 3**:
- **Step 4**:

Statistical test procedure

The standard procedure of the statistical test is the following.

- **Step 1**: Set the null hypothesis and alternative hypothesis. Also, fix the significance level α (usually 0.05 or 0.005).
- **Step 2**:
- **Step 3**:
- **Step 4**:

Statistical test procedure

The standard procedure of the statistical test is the following.

- **Step 1**: Set the null hypothesis and alternative hypothesis. Also, fix the significance level α (usually 0.05 or 0.005). For example, in the t-test, the null hypothesis is $\mu = \mu_0$, where μ is the unknown true mean and μ_0 is a value, which we decide. The alternative hypothesis is $\mu \neq \mu_0$

- **Step 2**:
- **Step 3**:

- **Step 4**:

Statistical test procedure

The standard procedure of the statistical test is the following.

- **Step 1**: Set the null hypothesis and alternative hypothesis. Also, fix the significance level α (usually 0.05 or 0.005). For example, in the t-test, the null hypothesis is $\mu = \mu_0$, where μ is the unknown true mean and μ_0 is a value, which we decide. The alternative hypothesis is $\mu \neq \mu_0$
- **Step 2**: Calculate the statistic.
- **Step 3**:
- **Step 4**:

Statistical test procedure

The standard procedure of the statistical test is the following.

- **Step 1**: Set the null hypothesis and alternative hypothesis. Also, fix the significance level α (usually 0.05 or 0.005). For example, in the t-test, the null hypothesis is $\mu = \mu_0$, where μ is the unknown true mean and μ_0 is a value, which we decide. The alternative hypothesis is $\mu \neq \mu_0$
- **Step 2**: Calculate the statistic. For example, in the t-test, calculate $t = \frac{\sqrt{m}\left(\overline{X} - \mu_0\right)}{s}$.
- **Step 3**:
- **Step 4**:

Statistical test procedure

The standard procedure of the statistical test is the following.

- **Step 1**: Set the null hypothesis and alternative hypothesis. Also, fix the significance level α (usually 0.05 or 0.005). For example, in the t-test, the null hypothesis is $\mu = \mu_0$, where μ is the unknown true mean and μ_0 is a value, which we decide. The alternative hypothesis is $\mu \neq \mu_0$
- **Step 2**: Calculate the statistic. For example, in the t-test, calculate $t = \frac{\sqrt{m}\left(\overline{X} - \mu_0\right)}{s}$.
- **Step 3**: Evaluate the p-value from the value of the statistic.
- **Step 4**:

Statistical test procedure

The standard procedure of the statistical test is the following.

- **Step 1**: Set the null hypothesis and alternative hypothesis. Also, fix the significance level α (usually 0.05 or 0.005). For example, in the t-test, the null hypothesis is $\mu = \mu_0$, where μ is the unknown true mean and μ_0 is a value, which we decide. The alternative hypothesis is $\mu \neq \mu_0$
- **Step 2**: Calculate the statistic. For example, in the t-test, calculate $t = \frac{\sqrt{m}\left(\overline{X} - \mu_0\right)}{s}$.
- **Step 3**: Evaluate the p-value from the value of the statistic. For example, in the t-test, we can evaluate p-value by referring to t-tables.
- **Step 4**:

Statistical test procedure

The standard procedure of the statistical test is the following.

- **Step 1**: Set the null hypothesis and alternative hypothesis. Also, fix the significance level α (usually 0.05 or 0.005). For example, in the t-test, the null hypothesis is $\mu = \mu_0$, where μ is the unknown true mean and μ_0 is a value, which we decide. The alternative hypothesis is $\mu \neq \mu_0$
- **Step 2**: Calculate the statistic. For example, in the t-test, calculate $t = \frac{\sqrt{m}\left(\overline{X} - \mu_0\right)}{s}$.
- **Step 3**: Evaluate the p-value from the value of the statistic. For example, in the t-test, we can evaluate p-value by referring to t-tables.
- **Step 4**: If $p < \alpha$, then we reject the null hypothesis and accept the alternative hypothesis. If $p \leq \alpha$, we can **neither reject nor accept a hypothesis**.

t-test example

Example

Our factory composed a component 8 times and the purity was $(95, 93, 94, 94, 92, 93, 91, 96)$.
Suppose that the purity of a natural one is 92% on average.
Are our factory's products better than natural ones on average?

t-test example

Example

Our factory composed a component 8 times and the purity was $(95, 93, 94, 94, 92, 93, 91, 96)$.
Suppose that the purity of a natural one is 92% on average.
Are our factory's products better than natural ones on average?

Step 1: Set the null hypothesis and alternative hypothesis. Also, fix the significance level α (usually 0.05 or 0.005).

The null hypothesis is $\mu = \mu_0 = 92$. The alternative hypothesis is $\mu \neq \mu_0 = 92$. Let's use the significance level $\alpha = 0.05$.

t-test example

Example

Our factory composed a component 8 times and the purity was $(95, 93, 94, 94, 92, 93, 91, 96)$.
Suppose that the purity of a natural one is 92% on average.
Are our factory's products better than natural ones on average?

Step 2: Calculate the t-statistic.

The sample mean and standard deviation are $\overline{X} = 93.5$ and $s \approx 1.60$.

The t-statistic is $t = \frac{\sqrt{m}\left(\overline{X} - \mu_0\right)}{s} \approx \frac{93.5-92}{\frac{1.6}{2\sqrt{2}}} = 2.65$.

t-test example

Example

Our factory composed a component 8 times and the purity was $(95, 93, 94, 94, 92, 93, 91, 96)$.
Suppose that the purity of a natural one is 92% on average.
Are our factory's products better than natural ones on average?

Step 3: Evaluate the p-value from the value of the t-statistic.

Here, under the null hypothesis, t follows the t-distribution with 7 degrees of freedom.

Then, if $t \approx 2.65$, the p-value is $p \approx 0.032$, according to an online calculator.

t-test example

Example

Our factory composed a component 8 times and the purity was $(95, 93, 94, 94, 92, 93, 91, 96)$.
Suppose that the purity of a natural one is 92% on average.
Are our factory's products better than natural ones on average?

Step 4: Conclude from the p-value.

Since $p \approx 0.032 < \alpha = 0.05$, we reject the null hypothesis and accept the alternative hypothesis.

Hence, we can **statistically conclude that our factory produces better components than natural ones.**

Outline

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

Statistical tests behave stochastically, so they 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.

Outline

Exercise (Statistical test 1)

Suppose that we apply a statistical test.
If the p-value of the test statistic is **lower** than the significance level, what of the following are the correct actions? Select all that apply.

- Accept the null hypothesis.
- Reject the null hypothesis.
- Accept the alternative hypothesis.
- Reject the alternative hypothesis.

Exercise (Statistical test 2)

Suppose that we apply a statistical test.
If the p-value of the test statistic is **higher** than the significance level, what of the following are the correct actions? Select all that apply.

- Accept the null hypothesis.
- Reject the null hypothesis.
- Accept the alternative hypothesis.
- Reject the alternative hypothesis.

Exercise (t-test 1)

Our factory composed a component 24 times and the purity was $(95, 93, 94, 94, 92, 93, 91, 96, 93, 95, 96, 91, 92, 93, 94, 94, 91, 95, 96, 93, 94, 93, 94, 92)$.
Suppose that the purity of a natural one is 92% on average.
Are our factory's products better than natural ones on average?
Set the significance level $\alpha = 0.05$.
You can use the fact that the t-distribution can be approximated by the standard normal distribution if the degree of freedom is larger than 20.
Note that $\Pr(Z \geq 2) \approx 0.025$, where Z is a random variable following the standard normal distribution.

Exercise (t-test 2)

Our factory composed a component 24 times and the purity was $(95, 93, 94, 94, 92, 93, 91, 96, 93, 95, 96, 91, 92, 93, 94, 94, 91, 95, 96, 93, 94, 93, 94, 92)$.
Suppose that the purity of a natural one is 93.25% on average.
Are our factory's products better than natural ones on average?
Set the significance level $\alpha = 0.05$.
You can use the fact that the t-distribution can be approximated by the standard normal distribution if the degree of freedom is larger than 20.
Note that $\Pr(Z \geq 2) \approx 0.025$, where Z is a random variable following the standard normal distribution.