Probability Theory

SUZUKI, Atsushi

The whole contents

© SUZUKI, Atsushi

Outline

1 Random variables

- Introduction: why do we learn random variables?
- Univariate discrete random variable
- Visualization of a distribution
- Summary statistics for a univariate random variable
- Expectation
- Median
- Variance and a function of a random variable
- Exercises

© SUZUKI, Atsushi

Outline

1 Random variables

- Introduction: why do we learn random variables?

Probability theory handles random events

Probability theory handles random ***events***, where the probability $\Pr(A) \in [0, 1]$ is defined for each event A. Here, an event is a subset of the ***sample space***, the set of all the possible outcomes. Each element in the sample space is called an ***elementary event***.

Example (Weather forecast)

Consider a weather forecast of 24 hours later. The sample space $S = \{\text{(It will be) sunny, cloudy, rainy, snowy}\}$. Suppose that the probability for each elementary event is given by

Event A	$\{\text{sunny}\}$	$\{\text{cloudy}\}$	$\{\text{rainy}\}$	$\{\text{sunny}\}$
The probability $\Pr(A)$	0.4	0.2	0.3	0.1

If an event A includes multiple elements in the sample space S, the probability $\Pr(A)$ is given by the sum of the probabilities of those elements. For example, $\Pr(\{\text{sunny, rainy}\}) = \Pr(\{\text{sunny}\}) + \Pr(\{\text{rainy}\}) = 0.4 + 0.3 = 0.7$.

© SUZUKI, Atsushi

Probability theory handles random events

Probability theory handles random ***events***, where the probability $\Pr(A) \in [0,1]$ is defined for each event A. Here, an event is a subset of the ***sample space***, the set of all the possible outcomes. Each element in the sample space is called an ***elementary event***.

Example (Weather forecast)

Consider a weather forecast of 24 hours later. The sample space $S = \{\text{(It will be) sunny, cloudy, rainy, snowy}\}$. Suppose that the probability for each elementary event is given by

Event A	$\{\text{sunny}\}$	$\{\text{cloudy}\}$	$\{\text{rainy}\}$	$\{\text{sunny}\}$
The probability $\Pr(A)$	0.4	0.2	0.3	0.1

In the above example, each event is a set of real phenomena, which we do not regard as a numeric value directly. However, in the following, **we always assume that each event is a set of numeric values**.

© SUZUKI, Atsushi

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

© SUZUKI, Atsushi

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

But **WHY** do we limit the discussion to RVs only, instead of considering general random events? The reasons are the following:

- RVs, i.e., numeric random events, are **all the random events we need to handle in computer science**, including AI, since a computer can only handle numeric values.
- If random events are RVs, i.e., numeric, we can discuss their random behaviors **quantitatively** based on the RVs' numeric values.

© SUZUKI, Atsushi

Learning outcomes

By the end of this section, 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

1. **Random variables**

   - Univariate discrete random variable

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.

© SUZUKI, Atsushi

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.

© SUZUKI, Atsushi

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.

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
$\Pr(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Table: Rolling an ideal six-sided dice

Probability mass function (PMF)

When we consider a univariate discrete random variable taking a value in a discrete set $\mathscr{X} = \{x_1, x_2, \ldots\} \subset \mathbb{R}$, 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
$P_X(x) := \Pr(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Table: 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***.

© SUZUKI, Atsushi

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
$P_X(x) := \Pr(X = x) = \frac{m_x}{m}$	0.15	0.10	0.15	0.25	0.30	0.05

Table: Exam result data points and the frequency.

Outline

1. **Random variables**

   Visualization of a distribution

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.

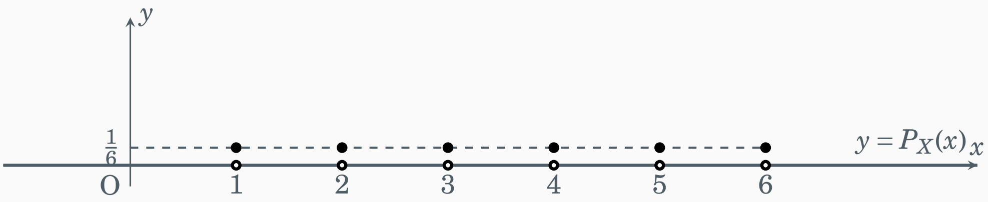
One way is to draw a graph of the PMF.

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
$P_X(x) := \Pr(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Table: The PMF of rolling an ideal six-sided dice



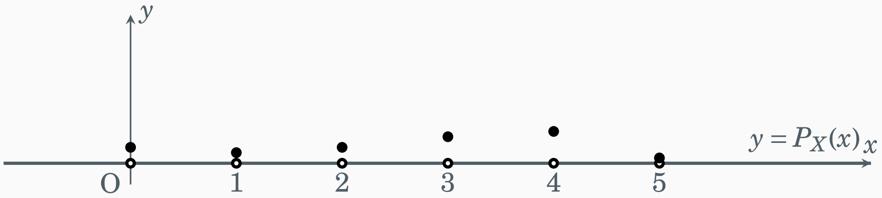
© SUZUKI, Atsushi

Example of a PMF graph: rolling an ideal dice

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

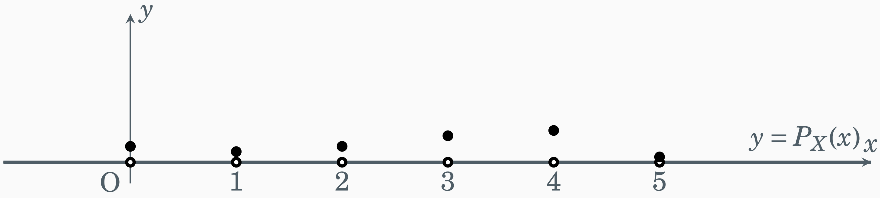
Score x	0	1	2	3	4	5
$P_X(x) := \Pr(X = x)$	0.15	0.10	0.15	0.25	0.30	0.05

Table: The PMF of the frequency of exam results



© SUZUKI, Atsushi

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., $\Pr(1.5 \le X \le 3.8)$.

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

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

© SUZUKI, Atsushi

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

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

Example of CDF: rolling an ideal dice

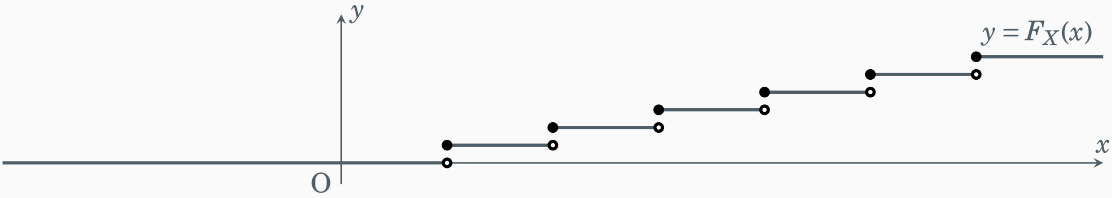
Suppose that we roll an ideal six-sided dice. The PMF is given as follows.

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

Table: 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.

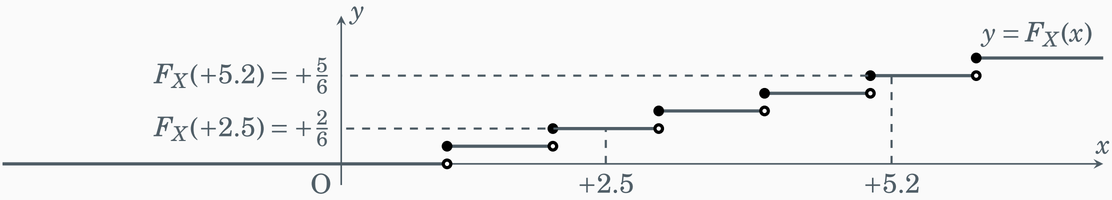


x	$(-\infty, 1)$	$[1, 2)$	$[2, 3)$	$[3, 4)$	$[4, 5)$	$[5, 6)$	$[6, +\infty)$
$F_X(x) := \Pr(X = x)$	0	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1

Table: 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.



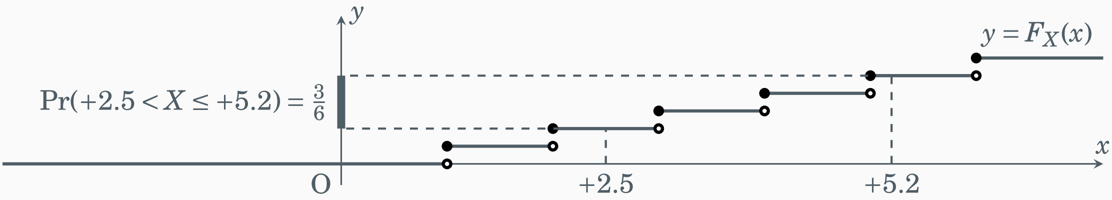
Using the CDF, we can calculate the probability of various events. For example,

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

© SUZUKI, Atsushi

Example of CDF: rolling an ideal dice

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

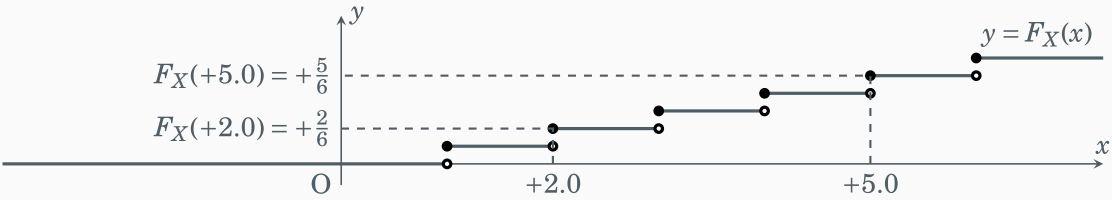


Using the CDF, we can calculate the probability of various events. For example,

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

Example of CDF: rolling an ideal dice

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

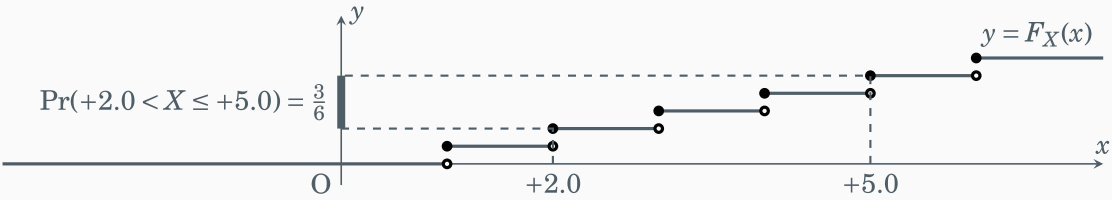


Using the CDF, we can calculate the probability of various events. For example,

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

Example of CDF: rolling an ideal dice

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

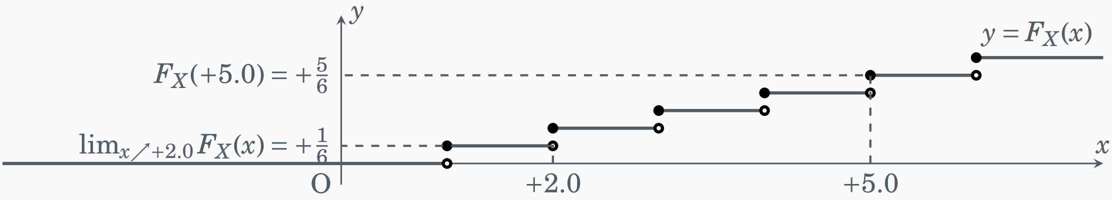


Using the CDF, we can calculate the probability of various events. For example,

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

Example of CDF: rolling an ideal dice

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

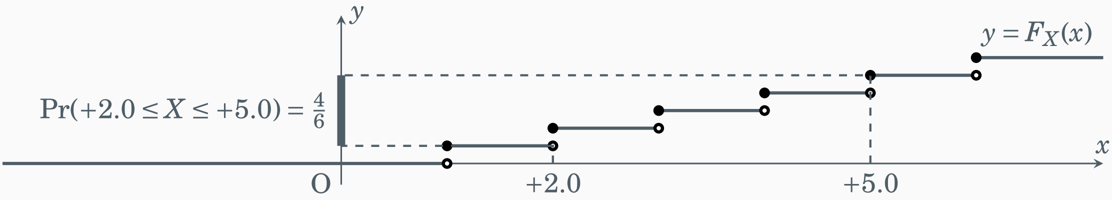


Using the CDF, we can calculate the probability of various events. For example,

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

Example of CDF: rolling an ideal dice

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

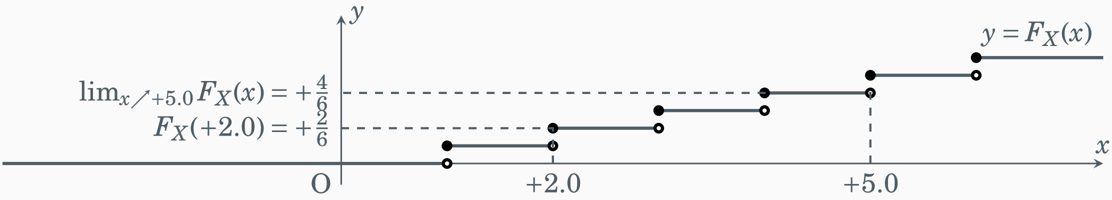


Using the CDF, we can calculate the probability of various events. For example,

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

Example of CDF: rolling an ideal dice

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

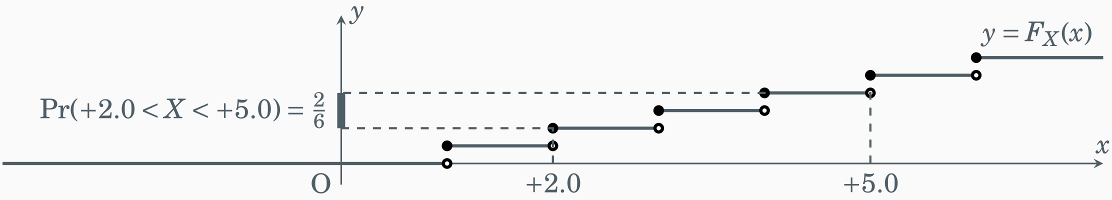


Using the CDF, we can calculate the probability of various events. For example,

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

Example of CDF: rolling an ideal dice

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

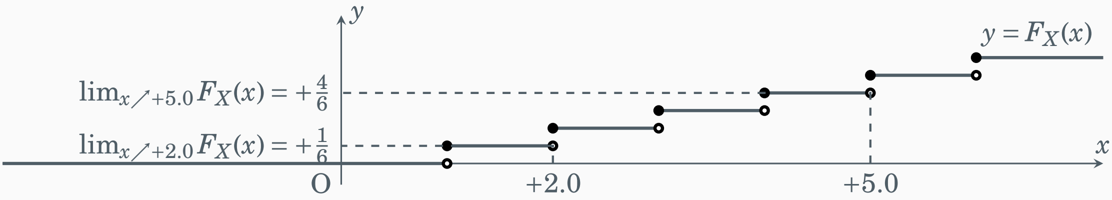


Using the CDF, we can calculate the probability of various events. For example,

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

Example of CDF: rolling an ideal dice

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

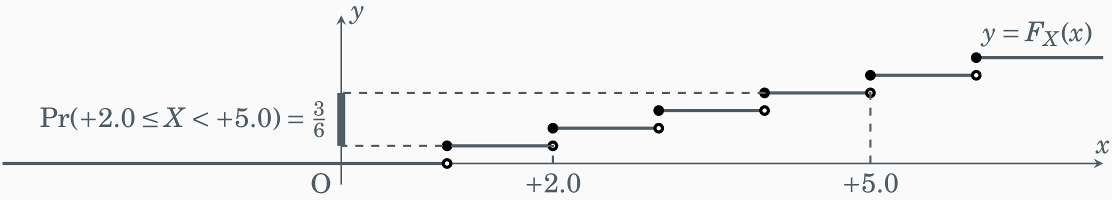


Using the CDF, we can calculate the probability of various events. For example,

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

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,

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

© SUZUKI, Atsushi

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

Table: 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

Table: 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

Table: The CDF of a student exam result frequency

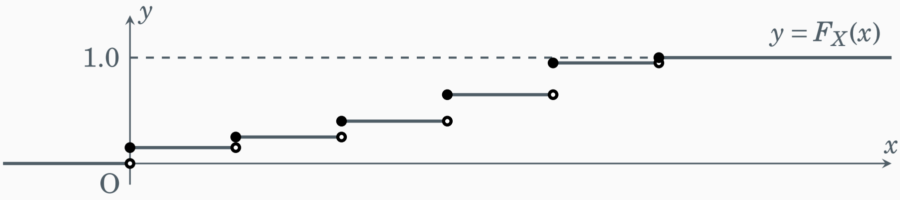
Example of CDF: student score frequency

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

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

Table: 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.

© SUZUKI, Atsushi

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

Theorem

Let $f : \mathbb{R} \to \mathbb{R}$ 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

1. **Random variables**

Summary statistics for a univariate random variable

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.

© SUZUKI, Atsushi

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.

© SUZUKI, Atsushi

Outline

1. **Random variables**

   - Expectation

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

Table: 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
$xP_X(x)$					

Table: 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				

Table: 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			

Table: 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		

Table: 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**:

© SUZUKI, Atsushi

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	$+0.10$	$+1.10$

Table: 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	$+0.10$	$+1.10$

Table: 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 $\mathbb{E}X$.

© SUZUKI, Atsushi

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	$+0.10$	$+1.10$

Table: 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.

© SUZUKI, Atsushi

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

Table: Exam result data points and the frequency.

© SUZUKI, Atsushi

Outline

1. **Random variables**

   Median

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.

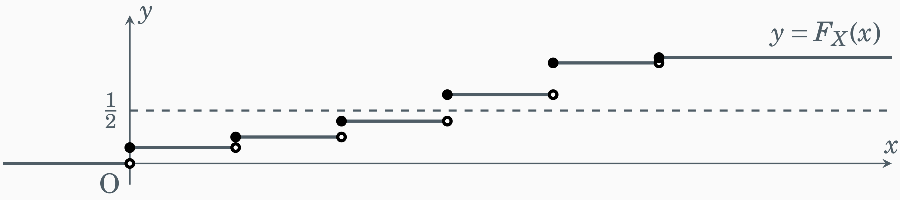
Roughly speaking, the median is defined so that the random variable is larger than the median in 50% probability and smaller than the median in 50% probability.

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 $y = \frac{1}{2}$.

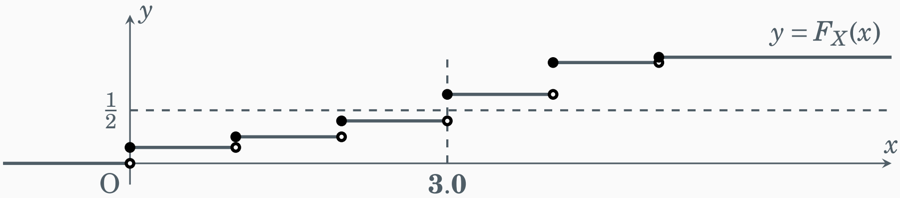


Median's idea

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 $y = \frac{1}{2}$.



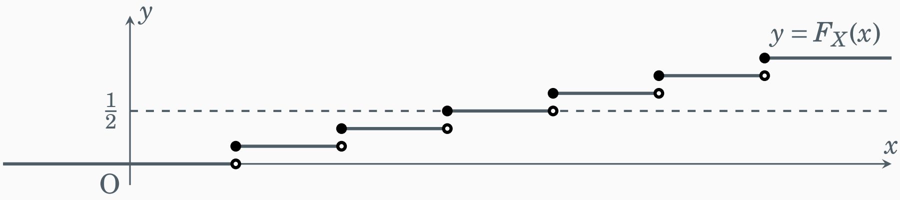
where the $y = F(x)$ crosses $y = \frac{1}{2}$

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.



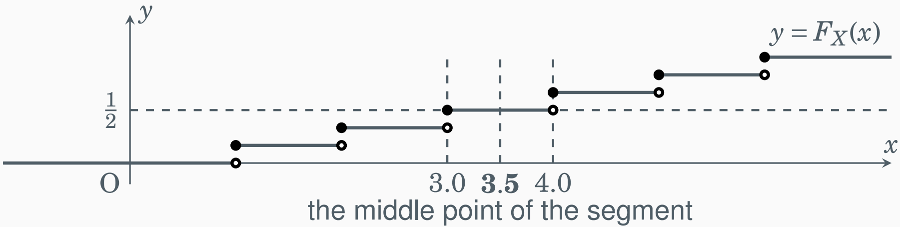
© SUZUKI, Atsushi

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.



© SUZUKI, Atsushi

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:

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

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

© SUZUKI, Atsushi

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:

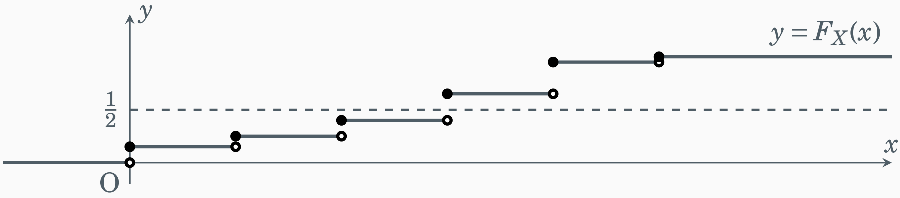
$$\begin{aligned} \underline{M} &:= \min \left\{ M \in \mathbb{R} \,\middle|\, \Pr(X \le M) \ge \frac{1}{2} \text{ and } \Pr(X \ge M) \ge \frac{1}{2} \right\}, \\ \overline{M} &:= \max \left\{ M \in \mathbb{R} \,\middle|\, \Pr(X \le M) \ge \frac{1}{2} \text{ and } \Pr(X \ge M) \ge \frac{1}{2} \right\}.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

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.

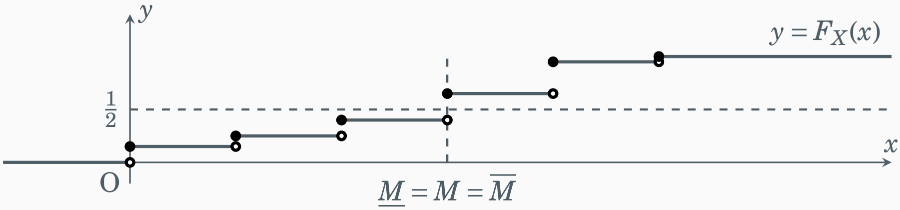
The definition of the median by the CDF graph

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



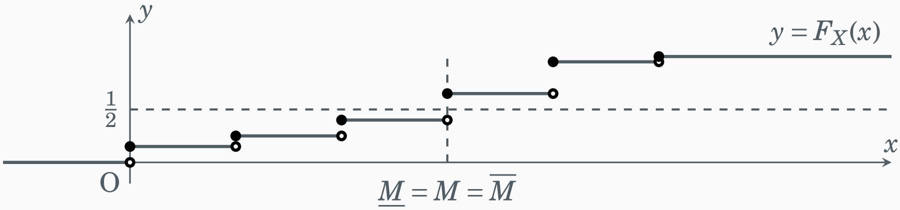
The definition of the median by the CDF graph

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

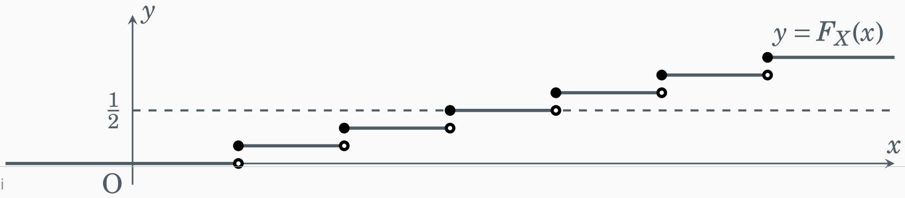


The definition of the median by the CDF graph

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



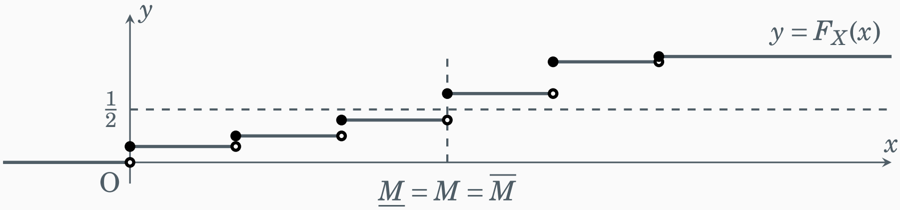
If the CDF graph has a horizontal segment on $y = \frac{1}{2}$,



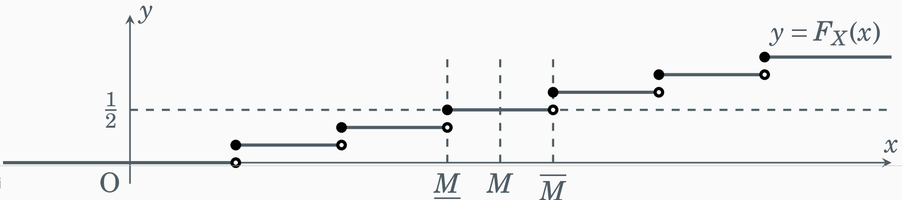
© SUZUKI, Atsushi

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 $y = \frac{1}{2}$,

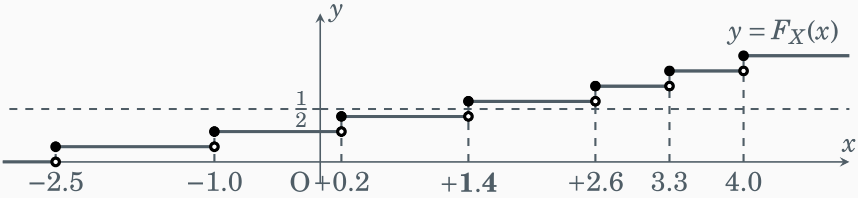


© SUZUKI, Atsushi

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.

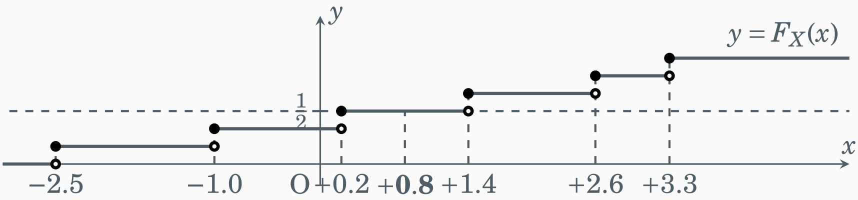
For example, if we have 7 sorted data points $(-2.5, -1.0, +0.2, +1.4, +2.6, +3.3, +4.0)$, then the median is the value of the 4th largest (or equivalently, the 4th smallest) data point, which is $+1.4$.



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 $\frac{0.2+1.4}{2} = 0.8$.



© SUZUKI, Atsushi

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

Table: 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

Table: 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

1. **Random variables**

   Variance and a function of a random variable

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 $(X-\mu)^2$?

© SUZUKI, Atsushi

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.

© SUZUKI, Atsushi

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

Table: 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

Table: 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.

© SUZUKI, Atsushi

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

Table: 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.

© SUZUKI, Atsushi

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

Table: 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.

© SUZUKI, Atsushi

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

Table: Example random function and its PMF.

To wrap up,

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

Table: The PMF of X^2.

© SUZUKI, Atsushi

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

Table: Example random function and its PMF.

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

© SUZUKI, Atsushi

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

Table: 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

Table: 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

Table: 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

Table: The PMF of $-3X+5$.

© SUZUKI, Atsushi

Proof: the linearity of the expectation

Proof.

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

□

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.

© SUZUKI, Atsushi

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 $\mathcal{X} \subset \mathbb{R}$. Also, suppose that $\mu := \mathbb{E}X$ and $P_X : \mathcal{X} \to [0,1]$ are its expectation and PMF, respectively.
The variance $\mathbb{V}[X]$ is given by

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

© SUZUKI, Atsushi

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. $\left(x - \mu_X\right)^2 P_X(x)$					

Table: 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. $\left(x - \mu_X\right)^2 P_X(x)$					

Table: 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. $\left(x - \mu_X\right)^2 P_X(x)$					

Table: 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. $(x - \mu_X)^2 P_X(x)$					

Table: 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				
Square deviation $(x-\mu_X)^2$	9.00				
Weighted sq. dev. $(x-\mu_X)^2 P_X(x)$	0.45				

Table: 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				

Table: 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 $(x - \mu_X)^2$	9.00	4.00			
Weighted sq. dev. $(x - \mu_X)^2 P_X(x)$	0.45				

Table: 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	4.00			
Weighted sq. dev. $\left(x - \mu_X\right)^2 P_X(x)$	0.45	0.40			

Table: 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**:

© SUZUKI, Atsushi

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

Table: 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 $x \in \mathcal{X}$.
- **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

Table: 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
  $\mathbb{V}[X] = \sum_{x \in \mathscr{X}} \left(x - \mu_X\right)^2 P_X(x) = 0.45 + 0.40 + 0.20 + 0.00 + 0.55$

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

Table: 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
  $\mathbb{V}[X] = \sum_{x \in \mathscr{X}} \left(x - \mu_X\right)^2 P_X(x) = 0.45 + 0.40 + 0.20 + 0.00 + 0.55 = 1.60.$

© SUZUKI, Atsushi

Another formula of the variance

The following formula is also useful.

Theorem

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

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

Proof.

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

□

© SUZUKI, Atsushi

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

Table: 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 $\mathbb{V}[-3X + 5]$.

© SUZUKI, Atsushi

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

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

© SUZUKI, Atsushi

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

Table: 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.

© SUZUKI, Atsushi

Outline

Exercise (Empirical distribution)

Consider a group of $m = 20$ students and their scores on a test. The scores are integers within the set $\mathscr{X} = \{0, 1, 2, 3, 4, 5\}$. For $x \in \mathscr{X}$, let m_x denote the number of students scoring x points. The results are given in the table below:

Score x	0	$+1$	$+2$	$+3$	$+4$	$+5$
Number of students m_x	3	2	3	5	6	1

Find the frequency (empirical distribution) of the data. Or, define a random variable X as the score when a student is chosen uniformly at random and calculate the probability mass function P_X.

Example answer:

The total number of data points is m, and for a value x, the number of data points taking the value x is m_x. Therefore, the probability mass function of the empirical distribution at x is given by $\frac{m_x}{m}$. Hence, $P_X(0) = \frac{3}{20} = 0.15$, $P_X(1) = \frac{2}{20} = 0.10$, $P_X(2) = \frac{3}{20} = 0.15$, $P_X(3) = \frac{5}{20} = 0.25$, $P_X(4) = \frac{6}{20} = 0.30$, $P_X(5) = \frac{1}{20} = 0.05$.

© SUZUKI, Atsushi

Exercise (Descriptive statistics)

Let X be a discrete random variable, with its probability mass function P_X given by the table below:

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

(1) Write down the cumulative distribution function (CDF) of X, F_X.
Additionally, evaluate the median of X, denoted as med_X.

(2) Evaluate the expectation, variance, and standard deviation of X, denoted as μ_X, σ_X^2, σ_X respectively.

(3) Define a new random variable $Z = 5X - 2$. Evaluate the expectation and variance of Z, denoted as μ_Z, σ_Z^2 respectively.

© SUZUKI, Atsushi

Example answer:

(1) The cumulative distribution function F_X is defined as $F_X(x) := \Pr(X \le x)$. For example, $F_X(+0.5) = \Pr(X \le +0.5) = P_X(-2) + P_X(-1) + P_X(0) = 0.35$. In the case of a discrete random variable, the CDF appears as a step function. Specifically, within intervals that carry no probability mass, the CDF remains constant. Whenever a discrete random variable X carries probability mass at $x = a$, meaning $P_X(a) > 0$, the value of the CDF F_X increases by $P_X(a)$ at $x = a$. Therefore,

$$F_X(x) = \begin{cases} 0 & \text{if } x < -2, \\ 0.05 & \text{if } -2 \le x < -1, \\ 0.15 & \text{if } -1 \le x < 0, \\ 0.35 & \text{if } 0 \le x < +1, \\ 0.45 & \text{if } +1 \le x < +2, \\ 1 & \text{if } x \ge +2. \end{cases}$$

Example answer:

(1, continued) The median of X, denoted as med_X, is the value of x where the graph of $y = F_X(x)$ crosses the horizontal line $y = \frac{1}{2}$. Precisely, if for some real number a, $\lim_{x \nearrow a} F_X(x) < 0.5$ and $F_X(a) > 0.5$, then a is the median of X. In this case, since $\lim_{x \nearrow +2} F_X(x) = 0.45$ and $F_X(+2) = 1$, the graph of $y = F_X(x)$ crosses the horizontal line $y = \frac{1}{2}$ at $x = 2$, making the median of X equal to $+2$.

© SUZUKI, Atsushi

Example answer:

(2) For expectation and variance, generally, if X is a discrete random variable with support $\mathscr{X}$ and probability mass function P_X, then the expectation μ_X, variance σ_X^2, and standard deviation σ_X are given by:

$$\mu_X = \sum_{x \in \mathscr{X}} x P_X(x)$$

,

$$\sigma_X^2 = \sum_{x \in \mathscr{X}} (x - \mu_X)^2 P_X(x) = \sum_{x \in \mathscr{X}} x^2 P_X(x) - (\mu_X)^2$$

,

$$\sigma_X = \sqrt{\sigma_X^2}$$

.

Thus, $\mu_X = (-2) \cdot 0.05 + (-1) \cdot 0.10 + 0 \cdot 0.20 + (+1) \cdot 0.10 + (+2) \cdot 0.55 = +1$, $\sigma_X^2 = \sum_{x \in \mathscr{X}} x^2 P_X(x) - (\mu_X)^2 = 1.6$, and $\sigma_X = \sqrt{1.6} \approx 1.264$.

© SUZUKI, Atsushi

Example answer:

(3) For a transformed random variable $Z = aX + b$, the general formulas for expectation and variance are $\mu_Z = a\mu_X + b$ and $\sigma_Z^2 = a^2\sigma_X^2$. Hence, in this scenario, $\mu_Z = 5\mu_X - 2 = +3$ and $\sigma_Z^2 = 5^2\sigma_X^2 = 40$. Alternatively, calculating the probability mass function of Z and then computing expectation and variance in the same manner as done for X is also valid. Note that the probability mass function P_Z of Z is given by the following table.

z	-12	-7	-2	$+3$	$+8$
$P_Z(z)$	0.05	0.10	0.20	0.10	0.55

Exercise (Coin toss)

Consider defining a discrete random variable X through a coin toss, where the coin is placed on a finger with one side up, flicked, and then observed which side lands facing up. The support of X is $\{-1, +1\}$, with $X = +1$ if the side that was initially up is also up after the coin lands, and $X = -1$ otherwise. The probability mass function P_X of X is given by the following table:

x	-1	$+1$
$P_X(x)$	0.492	0.508

Given this, calculate the expected value and median of X, denoted as μ_X and med_X respectively. It is known that coin tosses in the real world follow a probability distribution closely resembling the one described above.

Example answer:

For a discrete random variable X with set $\mathscr{X}$ and probability mass function P_X, the expected value μ_X is given by $\mu_X = \sum_{x \in \mathscr{X}} x P_X$. Thus, for this problem, $\mu_X = (-1) \cdot 0.492 + (+1) \cdot 0.508 = 0.016$.

If for some real number a, $\lim\limits_{x \nearrow a} F_X(x) < 0.5$ and $F_X(a) > 0.5$, then a is the median of X. Since $\lim\limits_{x \nearrow 1} F_X(x) = 0.492$ and $F_X(1) = 1$, the median of X is 1.

© SUZUKI, Atsushi

Exercise (Average Income)

A village of 999 people had everyone earning an annual income of 10,000 pounds. The village chief aimed to double the average annual income of the village's residents. The next day, the village chief invited a billionaire with an annual income of x pounds, who then became a resident of the village. While the incomes of the other 999 residents remained unchanged, the billionaire's relocation resulted in the village's average annual income rising to 20,000 pounds.

Calculate the billionaire's annual income x.

Additionally, find the median annual income of the village's residents before and after the billionaire's relocation.

Example answer:

The average of a set of data is the expected value of the frequency (empirical distribution) of the data, which can be calculated as the total sum of the data divided by the number of data points. Conversely, if the average and the number of data points are known, the total sum of the data can be found by multiplying these two values. Therefore, before the billionaire's relocation, the total annual income of the villagers was $999 \times 10000 = 9990000$ pounds, and after the relocation, it became $1000 \times 20000 = 20000000$ pounds. Hence, the billionaire's annual income is 10010000 pounds.

The median annual income of the villagers before the billionaire's relocation is the 500th value when sorting the incomes of 999 villagers, all of whom earn 10000 pounds, thus the median is naturally 10000 pounds. After the relocation, the median of the annual incomes of 1000 villagers is the average of the 500th and 501st values from the top, again all 999 others earn 10000 pounds, so the median remains 10000 pounds.

© SUZUKI, Atsushi

Outline

2 Multiple Random Variables

- Introduction: why are multiple random variables less trivial?
- Joint distribution
- Marginal distribution
- Conditional distribution
- Independence of random variables
- Summary statistics for multiple RVs and covariance
- Correlation
- Exercises

Outline

2. **Multiple Random Variables**
   - Introduction: why are multiple random variables less trivial?

Multiple random variables

There are many cases where we handle multiple random variables (multiple RVs) in real applications as follows.

Example

- The prices of multiple stocks.
- The pixels of an image taken in the real world.
- The values at each time frame in a wave file of a human speech.

Since we deal with many multiple RVs in many real applications, it is natural to discuss them. A naïve way to handle multiple RVs is to apply the theory for a univariate RV to each of the multiple RVs that we are interested in. Is it sufficient?

© SUZUKI, Atsushi

Multiple random variables

There are many cases where we handle multiple random variables (multiple RVs) in real applications as follows.

Example

- The prices of multiple stocks.
- The pixels of an image taken in the real world.
- The values at each time frame in a wave file of a human speech.

Since we deal with many multiple RVs in many real applications, it is natural to discuss them. A naïve way to handle multiple RVs is to apply the theory for a univariate RV to each of the multiple RVs that we are interested in. Is it sufficient?

The answer is **NO**. Specifically, when we consider multiple random variables, knowing each probability mass function (PMF) is not sufficient to know their stochastic behavior completely.

© SUZUKI, Atsushi

Knowing multiple RVs $\neq$ knowing multiple PMFs

The following example shows that knowing the PMF for each RV is not sufficient to completely understand the random behavior of multiple RVs.

Example

Let X and Y be discrete RVs, whose supports are both $\{-1, +1\}$. Also, suppose that we know that the PMFs P_X and P_Y are given by $P_X(-1) = P_X(+1) = P_Y(-1) = P_Y(+1) = 0.5$.
Now, we know the exact distribution of X and Y. Still, we do not know the behavior of X and Y completely. For example, the above information does not determine the probability $\Pr(X = -1 \wedge Y = -1)$, where $\wedge$ indicates the logical “and” operator.

To know the random behavior of multiple RVs, we need to know the distribution of the **pair** (X, Y), which is called the ***joint distribution*** of the random variables X and Y. How to describe the joint distribution is the starting point of this section.

© SUZUKI, Atsushi

Learning outcomes

By the end of this section, 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

2 Multiple Random Variables

Joint distribution

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

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

where $\wedge$ indicates the logical "and" operator.

Properties of a joint PMF

From the properties of a probability distribution, we can easily see that a joint PMF satisfies the following.

Theorem

Let X and Y be discrete RVs and the joint PMF be $P_{X,Y}$.
Then, we have the following.

- *The probability mass is nonnegative everywhere, i.e.,* $0 \leq P_{X,Y}(x,y) \leq 1$ *for all* $(x,y) \in \mathcal{X} \times \mathcal{Y}$.
- *The sum of the probability masses is one, i.e.,* $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x,y) = 1$.

Joint PMF example

Example

Let X and Y be the scores of a math test and a history test, respectively, where we uniform-randomly sample a student. In other words, X and Y are frequencies.
Then X and Y be discrete random variables. The joint PMF may look like the following.

		x			
		0	1	2	3
	0	0.16	0.04	0.02	0.06
y	1	0.18	0.04	0.04	0.16
	2	0.06	0.02	0.08	0.14

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

Joint PMF example

Example

Let X and Y be the scores of a math test and a history test, respectively, where we uniform-randomly sample a student. In other words, X and Y are frequencies.
Then X and Y be discrete random variables. The joint PMF may look like the following.

		x			
		0	1	2	3
	0	0.16	0.04	0.02	0.06
y	1	0.18	0.04	**0.04**	0.16
	2	0.06	0.02	0.08	0.14

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

Here, for example, $\Pr(X = 2 \wedge Y = 1) = P_{X,Y}(2,1) = \mathbf{0.04}$.

© SUZUKI, Atsushi

Joint PMF example

Example

Let X and Y be the scores of a math test and a history test, respectively, where we uniform-randomly sample a student. In other words, X and Y are frequencies.
Then X and Y be discrete random variables. The joint PMF may look like the following.

		x			
		0	1	2	3
	0	0.16	0.04	0.02	0.06
y	1	0.18	0.04	0.04	0.16
	2	0.06	0.02	0.08	0.14

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

In the above example, we can confirm that the probability mass is nonnegative everywhere, i.e., $0 \le P_{X,Y}(x,y) \le 1$ for all $(x,y) \in \mathscr{X} \times \mathscr{Y}$.

© SUZUKI, Atsushi

Joint PMF example

Example

Let X and Y be the scores of a math test and a history test, respectively, where we uniform-randomly sample a student. In other words, X and Y are frequencies.
Then X and Y be discrete random variables. The joint PMF may look like the following.

		x			
		0	1	2	3
	0	0.16	0.04	0.02	0.06
y	1	0.18	0.04	0.04	0.16
	2	0.06	0.02	0.08	0.14

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

Also, we can see that the sum of the probability masses is one, i.e., $\sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} P_{X,Y}(x,y) = 1$.

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

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

© SUZUKI, Atsushi

Outline

2 Multiple Random Variables

Marginal distribution

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	
$P_X(x)$						

Table: 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.

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

© SUZUKI, Atsushi

Marginal distribution example

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

Table: 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.

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

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				

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

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

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

© SUZUKI, Atsushi

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**			

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

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

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

© SUZUKI, Atsushi

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			

Table: 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.

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

© SUZUKI, Atsushi

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**		

Table: 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.

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

© SUZUKI, Atsushi

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		

Table: 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				
		0	1	2	3	$P_Y(y)$
	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**	

Table: 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.

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

Marginal distribution example

		x 0	1	2	3	$P_Y(y)$
	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	

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

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

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

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	

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

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

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

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	

Table: 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.

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

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	

Table: 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.

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

© SUZUKI, Atsushi

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	

Table: 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	

Table: 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

2 **Multiple Random Variables**

- Conditional distribution

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				$P_Y(y)$
		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	

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

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	

Table: 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 $P_{X|Y}(x|2)$.

© SUZUKI, Atsushi

Conditional probability calculation

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

Table: 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.,

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

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
$P_{X\|Y}(x\|y)$					

Table: Joint PMF and conditional PMF

For example,

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
$P_{X\|Y}(x\|y)$	?				

Table: Joint PMF and conditional PMF

For example,

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

© SUZUKI, Atsushi

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**				

Table: 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	?			

Table: Joint PMF and conditional PMF

For example,

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

© SUZUKI, Atsushi

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**			

Table: 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 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	?		

Table: Joint PMF and conditional PMF

For example,

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

© SUZUKI, Atsushi

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**		

Table: Joint PMF and conditional PMF

For example,

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

© SUZUKI, Atsushi

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	?	

Table: 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 = 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**	

Table: Joint PMF and conditional PMF

For example,

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

© SUZUKI, Atsushi

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	
$P_X(x)$	0.40	0.10	0.20	0.30	

Table: 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.

© SUZUKI, Atsushi

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**.

© SUZUKI, Atsushi

Conditional probability calculation from joint PMF

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

2 **Multiple Random Variables**

Independence of random variables

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.

© SUZUKI, Atsushi

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	0.20	0.05	0.10	0.15	0.50
	2	0.12	0.03	0.06	0.09	0.30
$P_X(x)$		0.40	0.10	0.20	0.30	

Table: 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	$P_Y(y)$
y	0	**0.08**	0.02	0.04	0.06	0.20
	1	0.20	0.05	0.10	0.15	0.50
	2	0.12	0.03	0.06	0.09	0.30
$P_X(x)$		0.40	0.10	0.20	0.30	

Table: 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				$P_Y(y)$
		0	1	2	3	
	0	0.08	0.02	0.04	0.06	0.20
y	1	0.20	0.05	**0.10**	0.15	0.50
	2	0.12	0.03	0.06	0.09	0.30
$P_X(x)$		0.40	0.10	0.20	0.30	

Table: 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 $P_X(2)P_Y(1) = 0.20 \times 0.50$.

2 Multiple Random Variables

Summary statistics for multiple RVs and covariance

© SUZUKI, Atsushi

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.

© SUZUKI, Atsushi

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.

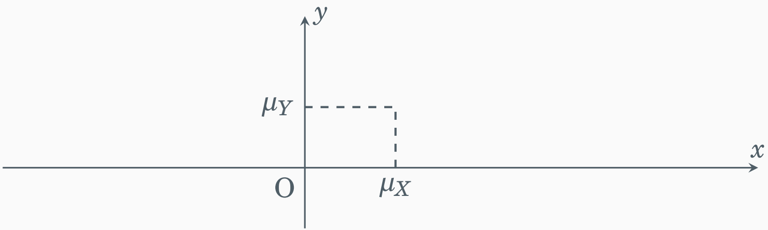


Figure:

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.

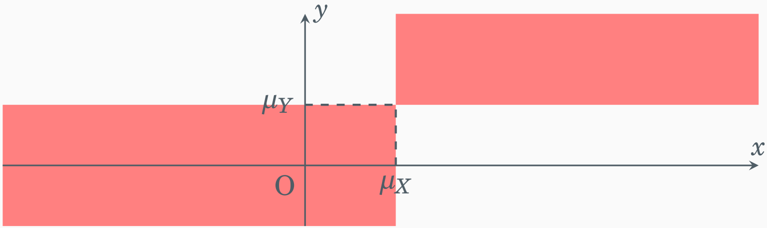


Figure: The area where $(X - \mu_X)(Y - \mu_Y)$ takes a positive value.

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.

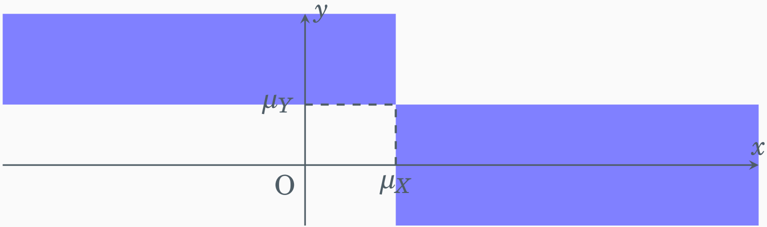


Figure: The area where $(X - \mu_X)(Y - \mu_Y)$ takes a negative value.

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 $(X-\mu_X)(Y-\mu_Y)$?

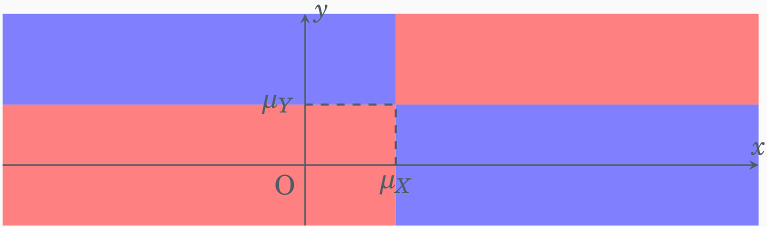


Figure:

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.

© SUZUKI, Atsushi

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 $\mathcal{X}$ and $\mathcal{Y}$, 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 $\{f(x,y) | x \in \mathcal{X}, y \in \mathcal{Y}\}$,*
- *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*

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

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

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

© SUZUKI, Atsushi

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

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

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

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

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	

Table: 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	$+1$	$P_Y(y)$
	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	

Table: 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 $y \in \mathscr{Y}$.

© SUZUKI, Atsushi

Example of covariance calculation

Example

		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	

Table: 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	

Table: 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	

Table: 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 $(x,y)\in\mathcal{X}\times\mathcal{Y}$ and take the sum.

  In the above example, we have $\mathrm{Cov}(X,Y)=\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}}(x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$
  $=(-\mathbf{0.5})\cdot(-\mathbf{1})\cdot\mathbf{0.25}$

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	

Table: 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	

Table: 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 $\mathrm{Cov}(X,Y) = \sum_{(x,y)\in\mathscr{X}\times\mathscr{Y}} (x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$
  $= (-0.5)\cdot(-1)\cdot 0.25 + (+0.5)\cdot(-1)\cdot 0.00 + (\mathbf{-0.5})\cdot\mathbf{0}\cdot\mathbf{0.25}$

© SUZUKI, Atsushi

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	

Table: 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 $\mathrm{Cov}(X,Y)=\sum_{(x,y)\in\mathscr{X}\times\mathscr{Y}}(x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$
  $=(-0.5)\cdot(-1)\cdot 0.25+(+0.5)\cdot(-1)\cdot 0.00+(-0.5)\cdot 0\cdot 0.25+\cdots+\mathbf{0.5}\cdot\mathbf{1}\cdot\mathbf{0.25}$

© SUZUKI, Atsushi

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	

Table: 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 $\mathrm{Cov}(X,Y)=\sum_{(x,y)\in\mathscr{X}\times\mathscr{Y}}(x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)$
  $=(-0.5)\cdot(-1)\cdot 0.25+(+0.5)\cdot(-1)\cdot 0.00+(-0.5)\cdot 0\cdot 0.25+\cdots+0.5\cdot 1\cdot 0.25=0.25.$

© SUZUKI, Atsushi

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 $X_1, X_2, \ldots, X_m$.

© SUZUKI, Atsushi

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	+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	

Table: 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 $\begin{bmatrix} \mathrm{Cov}(X,X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Cov}(Y,Y) \end{bmatrix} = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.5 \end{bmatrix}$.

Outline

2 **Multiple Random Variables**

- Correlation

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.

© SUZUKI, Atsushi

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

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

The correlation coefficient is often denoted by ρ.

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

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

© SUZUKI, Atsushi

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	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	

Table: 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

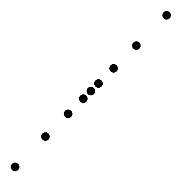


Figure: $\rho = 1.0$

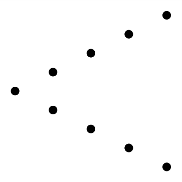


Figure: $\rho = 0.0$

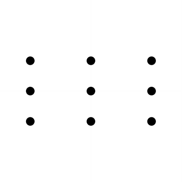


Figure: $\rho = 0.0$

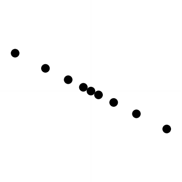


Figure: $\rho = -1.0$

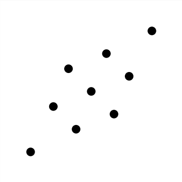


Figure: $\rho = 0.735$

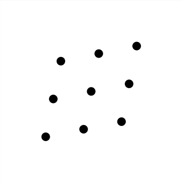


Figure: $\rho = 0.385$

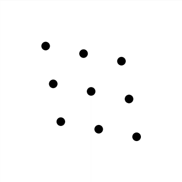


Figure: $\rho = -0.385$

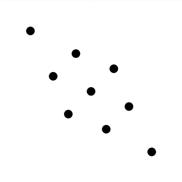


Figure: $\rho = -0.735$

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	

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

Table: 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.

© SUZUKI, Atsushi

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

© SUZUKI, Atsushi

Outline

2 **Multiple Random Variables**

Exercise (Joint PMF)

Consider two discrete random variables X and Y with supports $\mathscr{X} = \{0,1,2,3\}$ and $\mathscr{Y} = \{0,1,2\}$, respectively. Their joint probability mass function (joint PMF) $P_{X,Y}$ is given by the following table:

$P_{X,Y}(x,y)$	$x=0$	$x=1$	$x=2$	$x=3$
$y=0$	0.08	0.02	0.04	0.06
$y=1$	0.20	0.05	0.10	0.15
$y=2$	0.12	0.03	0.06	0.09

For example, the value 0.09 located in the column under $x=3$ and in the row for $y=2$ means $P_{X,Y}(3,2)=0.09$. Answer the following questions:

(1) Calculate the marginal PMFs P_X and P_Y.

(2) Let the conditional probability mass function (PMF) of X given Y and that of Y given X be denoted as $P_{X|Y}$ and $P_{Y|X}$, respectively. $P_{X|Y}(x|y)$ represents the probability that $X=x$ given the condition $Y=y$. Calculate the values of $P_{X|Y}(x|2)$ for all $x \in \mathscr{X}$ and $P_{Y|X}(y|1)$ for all $y \in \mathscr{Y}$.

(3) Determine whether the random variables X and Y are mutually independent or not.

© SUZUKI, Atsushi

Example answer:

(1) The marginal probability mass function (PMF) P_X for X is defined for each x in $\mathscr{X}$ as $P_X(x) := \Pr(X = x)$, which is the sum of the corresponding joint PMF values. Specifically, $P_X(x) := \Pr(X = x) = \sum_{y \in \mathscr{Y}} \Pr(X = x \land Y = y) = \sum_{y \in \mathscr{Y}} P_{X,Y}(x, y)$. Similarly, the marginal PMF P_Y for Y is defined for each y in $\mathscr{Y}$ as $P_Y(y) := \Pr(Y = y)$, given by $P_Y(y) = \sum_{x \in \mathscr{X}} P_{X,Y}(x, y)$.

In this problem:

$P_X(0) = P_{X,Y}(0,0) + P_{X,Y}(0,1) + P_{X,Y}(0,2) = 0.08 + 0.20 + 0.12 = 0.40,$
$P_X(1) = P_{X,Y}(1,0) + P_{X,Y}(1,1) + P_{X,Y}(1,2) = 0.02 + 0.05 + 0.03 = 0.10,$
$P_X(2) = P_{X,Y}(2,0) + P_{X,Y}(2,1) + P_{X,Y}(2,2) = 0.04 + 0.10 + 0.06 = 0.20,$
$P_X(3) = P_{X,Y}(3,0) + P_{X,Y}(3,1) + P_{X,Y}(3,2) = 0.06 + 0.15 + 0.09 = 0.30,$
$P_Y(0) = P_{X,Y}(0,0) + P_{X,Y}(1,0) + P_{X,Y}(2,0) + P_{X,Y}(3,0) = 0.08 + 0.02 + 0.04 + 0.06 = 0.20,$
$P_Y(1) = P_{X,Y}(0,1) + P_{X,Y}(1,1) + P_{X,Y}(2,1) + P_{X,Y}(3,1) = 0.20 + 0.05 + 0.10 + 0.15 = 0.50,$
$P_Y(2) = P_{X,Y}(0,2) + P_{X,Y}(1,2) + P_{X,Y}(2,2) + P_{X,Y}(3,2) = 0.12 + 0.03 + 0.06 + 0.09 = 0.30.$

© SUZUKI, Atsushi

Example answer:

(2) The value of the conditional PMF $P_{X|Y}$, when Y is given, is obtained by normalizing the corresponding joint PMF values such that the total sums to 1. Specifically, for each x in $\mathscr{X}$ and y in $\mathscr{Y}$, $P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}$. Similarly, the value of the conditional PMF $P_{X|Y}$ is given by normalizing the corresponding joint PMF values. Specifically, for each x in $\mathscr{X}$ and y in $\mathscr{Y}$, $P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}$.

In this problem:

$$P_{X|Y}(0|2) = \frac{P_{X,Y}(0,2)}{P_Y(2)} = \frac{0.12}{0.30} = 0.40,$$
$$P_{X|Y}(1|2) = \frac{P_{X,Y}(1,2)}{P_Y(2)} = \frac{0.03}{0.30} = 0.10,$$
$$P_{X|Y}(2|2) = \frac{P_{X,Y}(2,2)}{P_Y(2)} = \frac{0.06}{0.30} = 0.20,$$
$$P_{X|Y}(3|2) = \frac{P_{X,Y}(3,2)}{P_Y(2)} = \frac{0.09}{0.30} = 0.30,$$
$$P_{Y|X}(0|1) = \frac{P_{X,Y}(1,0)}{P_X(1)} = \frac{0.02}{0.10} = 0.20,$$
$$P_{Y|X}(1|1) = \frac{P_{X,Y}(1,1)}{P_X(1)} = \frac{0.05}{0.10} = 0.50,$$
$$P_{Y|X}(2|1) = \frac{P_{X,Y}(1,2)}{P_X(1)} = \frac{0.03}{0.10} = 0.30.$$

Example answer:

(3) Generally, for discrete random variables X and Y with supports $\mathscr{X}$ and $\mathscr{Y}$, and their respective marginal PMFs P_X and P_Y, along with a given joint PMF $P_{X,Y}$, a necessary and sufficient condition for X and Y to be mutually independent is that for any $(x,y) \in \mathscr{X} \times \mathscr{Y}$, $P_{X,Y}(x,y) = P_X(x)P_Y(y)$ must hold. Thus, verifying that $P_{X,Y}(x,y) = P_X(x)P_Y(y)$ for all pairs (x,y) confirms that X and Y are mutually independent. Conversely, if there exists any pair (x,y) for which $P_{X,Y}(x,y) \neq P_X(x)P_Y(y)$, then X and Y are not independent.

In this case, all pairs (x,y) in $\mathscr{X} \times \mathscr{Y}$ satisfy $P_{X,Y}(x,y) = P_X(x)P_Y(y)$. For instance, $P_{X,Y}(1,2) = 0.03$, which matches $P_X(1)P_Y(2) = 0.10 \cdot 0.30$, and the same holds true for all pairs (x,y) in $\mathscr{X} \times \mathscr{Y}$. Therefore, the random variables X and Y are mutually independent.

Exercise (Covariance and Correlation Coefficient)

Consider two discrete random variables X and Y with supports $\mathscr{X} = \{0, 1\}$ and $\mathscr{Y} = \{0, 1, 2\}$ respectively. The joint probability mass function (joint PMF) $P_{X,Y}$ is given by the following table:

$P_{X,Y}(x,y)$	$x = 0$	$x = 1$
$y = 0$	0.25	0.00
$y = 1$	0.25	0.25
$y = 2$	0.00	0.25

Evaluate the expected values of X and Y, denoted by μ_X and μ_Y respectively, the covariances $\mathrm{Cov}(X,X)$, $\mathrm{Cov}(X,Y)$, and $\mathrm{Cov}(Y,Y)$, and the correlation coefficient between X and Y, denoted by $\rho_{X,Y}$.

© SUZUKI, Atsushi

Example answer:

For this problem, the marginal probability mass functions P_X and P_Y for X and Y respectively are given by

x	0	1
$P_X(x)$	0.5	0.5

and

y	0	1	2
$P_Y(y)$	0.25	0.5	0.25

Thus, $\mu_X = 0 \cdot 0.5 + 1 \cdot 0.5 = 0.5$, $\mu_Y = 0 \cdot 0.25 + 1 \cdot 0.5 + 2 \cdot 0.25 = 1$. The covariance $\mathrm{Cov}(X,X)$ is equal to X's variance $\sigma_X^2 = \sum_{x \in \mathscr{X}} (x - \mu_X)^2 P_X(x) = \sum_{x \in \mathscr{X}} x^2 P_X(x) - (\mu_X)^2$. Therefore, in this problem, $\mathrm{Cov}(X,X) = 0^2 \cdot 0.5 + 1^2 \cdot 0.5 - 0.5^2 = 0.25$. Similarly, for Y, $\mathrm{Cov}(Y,Y) = 0^2 \cdot 0.25 + 1^2 \cdot 0.5 + 2^2 \cdot 0.25 - 1^2 = 0.5$.

© SUZUKI, Atsushi

Example answer (continued):

The covariance between X and Y, $\mathrm{Cov}(X,Y)$, can be calculated using the joint probability mass function $P_{X,Y}$, where $\mathrm{Cov}(X,Y) = \sum_{(x,y)\in\mathscr{X}\times\mathscr{Y}}(x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y) = \sum_{(x,y)\in\mathscr{X}\times\mathscr{Y}} xyP_{X,Y}(x,y) - \mu_X\mu_Y$. For this problem, $\mathrm{Cov}(X,Y) = 0\cdot 0\cdot 0.25 + 0\cdot 1\cdot 0.25 + 0\cdot 2\cdot 0.00 + 1\cdot 0\cdot 0.00 + 1\cdot 1\cdot 0.25 + 1\cdot 2\cdot 0.25 - 0.5\cdot 1 = 0.25$.

The correlation coefficient $\rho_{X,Y}$ is given by $\frac{\mathrm{Cov}(X,Y)}{\sqrt{\sigma_X^2}\sqrt{\sigma_Y^2}}$, yielding $\rho_{X,Y} = \frac{0.25}{\sqrt{0.25}\sqrt{0.5}} = \frac{1}{\sqrt{2}} \approx \frac{1}{1.414} \approx 0.707$.

Exercise

Consider discrete random variables X and Y, both with expectation and variance. Which of the following statements is correct? Select one option.

- The correlation coefficient between X and Y being 0 is a necessary and sufficient condition for X and Y to be independent.
- * The correlation coefficient between X and Y being 0 is a necessary condition for X and Y to be independent, but not a sufficient condition.
- The correlation coefficient between X and Y being 0 is a sufficient condition for X and Y to be independent, but not a necessary condition.
- The correlation coefficient between X and Y being 0 is neither a necessary condition nor a sufficient condition for X and Y to be independent.

Note: For conditions P, Q, if $P \implies Q$, meaning "if P, then Q" holds, then P is called a sufficient condition for Q, and Q is called a necessary condition for P.

Example answer:

First, let's clarify the definitions of the correlation coefficient and independence for discrete random variables. For discrete random variables X and Y, let $\mu_X := \mathbb{E}[X]$ and $\mu_Y := \mathbb{E}[Y]$ be the expected values of X and Y respectively, and $\sigma_X^2 := \mathbb{E}[(X-\mu_X)^2]$ and $\sigma_Y^2 := \mathbb{E}[(Y-\mu_Y)^2]$ be the variances of X and Y respectively. The covariance of X and Y, $\mathrm{Cov}(X,Y)$, is defined as $\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$. The correlation coefficient $\rho_{X,Y}$ is then defined as $\rho_{X,Y} := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\sigma_X^2}\sqrt{\sigma_Y^2}}$.

Furthermore, let $P_{X,Y}(x,y) := \Pr(X = x \wedge Y = y)$ be the joint probability mass function and $P_X(x) := \Pr(X = x)$ and $P_Y(y) := \Pr(Y = y)$ be the marginal probability mass functions for X and Y respectively. Independence between X and Y is defined as, for any x and y, $P_{X,Y}(x,y) = P_X(x)P_Y(y)$ holding true.

Example answer (continued):

In this problem, the statement "If the correlation coefficient between X and Y is 0, then X and Y are independent" does not hold. For example, in a case where $\Pr(X=0, Y=-1) = \Pr(X=0, Y=+1) = 0.25, \Pr(X=1, Y=0) = 0.5$, having a correlation coefficient of 0 does not imply that X and Y are independent. Therefore, the correlation coefficient being 0 is not a sufficient condition for X and Y to be independent.

On the other hand, if X and Y are independent, then their correlation coefficient is always 0. This can be proved as follows: For discrete random variables X and Y with supports $\mathscr{X}$ and $\mathscr{Y}$ respectively, let their expected values be μ_X and μ_Y, and their joint probability mass function be denoted by $P_{X,Y}$. Then, the covariance can be expressed as $\mathrm{Cov}(X, Y) = \sum_{x\in\mathscr{X}} \sum_{y\in\mathscr{Y}} xyP_{X,Y}(x,y) - \mu_X\mu_Y$. If X and Y are independent, then it always holds that $P_{X,Y}(x,y) = P_X(x)P_Y(y)$, so the first term becomes $\sum_{x\in\mathscr{X}} \sum_{y\in\mathscr{Y}} xyP_{X,Y}(x,y) = \sum_{x\in\mathscr{X}} \sum_{y\in\mathscr{Y}} xP_X(x)yP_Y(y) = \sum_{x\in\mathscr{X}} xP_X(x) \sum_{y\in\mathscr{Y}} yP_Y(y) = \mu_X\mu_Y$. Therefore, it follows that $\mathrm{Cov}(X, Y) = \mu_X\mu_Y - \mu_X\mu_Y = 0$. Thus, the correlation coefficient being 0 is a necessary condition for X and Y to be independent.

© SUZUKI, Atsushi

Outline

3 Continuous Random Variables

Outline

3 Continuous Random Variables

- Introduction: why are continuous random variables less trivial?

Continuous random variables in real AI applications

A discrete RV can take only limited values. However, many real-world phenomena are represented as random variables which can take any real value in a continuous section.

- Inflation rate (economics),
- Position of a vehicle,
- The brightness of scenery,
- The intensity of an acoustic signal,
- Density of air pollution.

Hence, when we want to analyze those phenomena using probability theory, we cannot always use mathematical tools to handle discrete RVs.

For example, those random variables typically have **no probability mass function (PMF)**.

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 an infinite number of real numbers in $[0,1]$, so $\Pr(X \in [0,1])$ is infinity. It contradicts $\Pr(X \in [0,1]) = 1$. □

© SUZUKI, Atsushi

A random variable may not have a PMF.

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

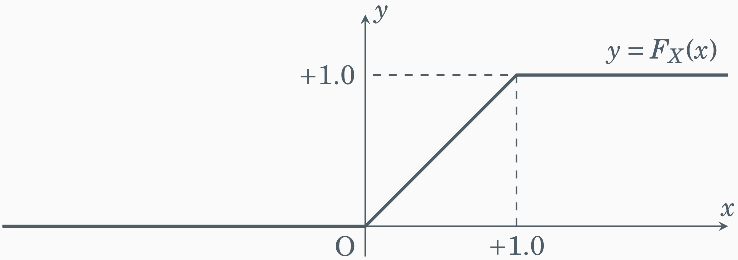
This random variable have nowhere probability mass, i.e., $\Pr(X = x) = 0$. for any $X \in \mathbb{R}$.

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)

Example of CDF for a non-discrete RV

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

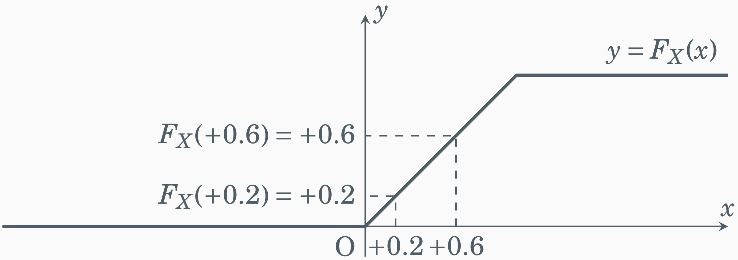


$$F_X(x) = \begin{cases} 0 & \text{if } x \le 0, \\ x & \text{if } 0 \le x \le 1, \\ 1 & \text{if } x \ge 1. \end{cases} \tag{37}$$

© SUZUKI, Atsushi

Example of CDF for a non-discrete RV

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

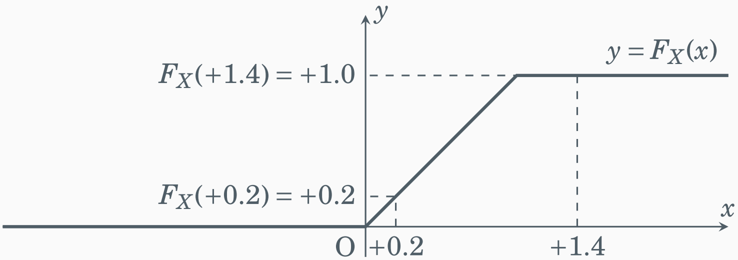


Using the CDF, we can calculate the probability of various events. For example,

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

Example of CDF for a non-discrete RV

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

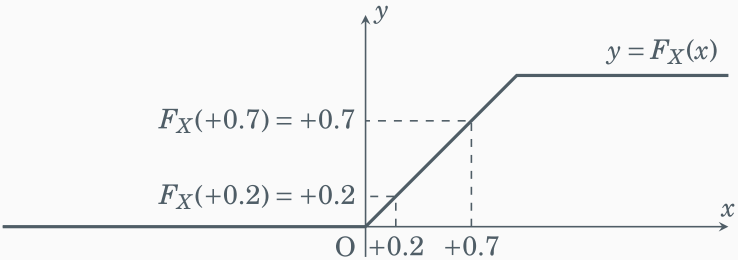


Using the CDF, we can calculate the probability of various events. For example,

$$\begin{aligned}\Pr(0.2 < X \leq 1.4) &= \Pr(X \leq 1.4) - \Pr(X \leq 0.2) \\ &= F_X(1.4) - F_X(0.2) \\ &= 1.0 - 0.2 = 0.8.\end{aligned} \tag{37}$$

Example of CDF for a non-discrete RV

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,

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

© SUZUKI, Atsushi

Why are we not satisfied with the CDF?

However, the CDF is not always welcomed. It is because

- 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.

© SUZUKI, Atsushi

Learning outcomes

By the end of this section, 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.

Notation: sections

In the following, $\mathbb{R}$, $\mathbb{R}_{\geq 0}$, and $\mathbb{R}_{>0}$ are the sets of real numbers, nonnegative real numbers, and positive real numbers, respectively.

Let a and b be real values. By $[a,b]$, (a,b), we denote the closed and open sections defined by

- $[a,b] = \{x \in \mathbb{R} | a \leq x \leq b\}$,
- $(a,b) = \{x \in \mathbb{R} | a < x < b\}$,

respectively. Likewise, by $(a,b]$ and $[a,b)$, we denote the semi-open sets defined by

- $(a,b] = \{x \in \mathbb{R} | a < x \leq b\}$,
- $[a,b) = \{x \in \mathbb{R} | a \leq x < b\}$,

Notation: Napier's constant and the exponential function

The real number constant e, called ***Napier's constant*** or ***Euler's number***, is defined by $\mathrm{e} := \lim_{n \to +\infty} \left(1 + \frac{1}{n}\right)^n$. Note that $\mathrm{e} = 2.718281828...$ and is the only real value that satisfies $\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{e}^x = \mathrm{e}^x$.

We define the ***(natural) exponential function*** $\exp : \mathbb{R} \to \mathbb{R}_{>0}$ by $\exp(x) = \mathrm{e}^x$.

© SUZUKI, Atsushi

Outline

3 **Continuous Random Variables**

- Probability density function

Idea of the probability density function

As we have seen in the case of the uniform distribution in the section $[0, 1]$, the probability $\Pr(X = c)$ might be zero for a real value c in many cases. In this case, we cannot say which values the RV tend to take more frequently than others.

Idea of the probability density function

As we have seen in the case of the uniform distribution in the section $[0,1]$, the probability $\Pr(X=c)$ might be zero for a real value c in many cases. In this case, we cannot say which values the RV tend to take more frequently than others.

Hence, we evaluate the probability of the RV taking a value **in a section**. For example, instead of evaluating $\Pr(X=c)$, we evaluate the probability $\Pr(a<X\leq b)$ for real values a,b around c such that $a<b$. If the probability is high and the section length $b-a$ is short, we can say that the RV X takes a value around c frequently.

© SUZUKI, Atsushi

Idea of the probability density function

As we have seen in the case of the uniform distribution in the section $[0,1]$, the probability $\Pr(X=c)$ might be zero for a real value c in many cases. In this case, we cannot say which values the RV tend to take more frequently than others.

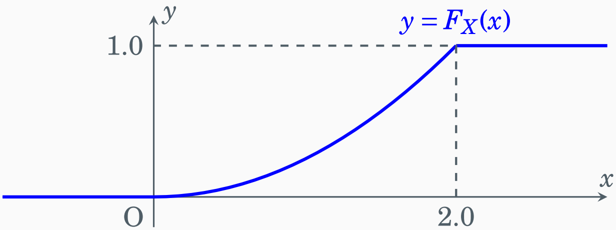
Hence, we evaluate the probability of the RV taking a value **in a section**. For example, instead of evaluating $\Pr(X=c)$, we evaluate the probability $\Pr(a<X\le b)$ for real values a,b around c such that $a<b$. If the probability is high and the section length $b-a$ is short, we can say that the RV X takes a value around c frequently.

So, we can regard the probability per the section length as the ***density*** of the probability distribution of the RV X around the section. A high density around a value c indicates that the X tends to take a value around c.

Based on the above idea, we can formulate the ***probability density function (PDF)*** from the cumulative distribution function (CDF) as follows.

© SUZUKI, Atsushi

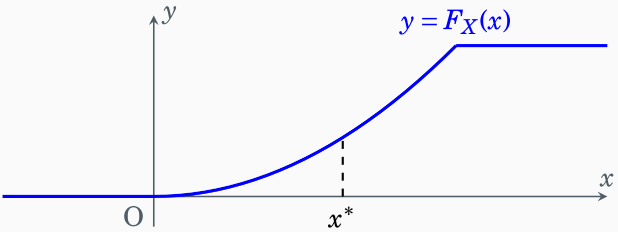
CDF to the density.



The CDF of a RV X.

Figure:

CDF to the density.



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

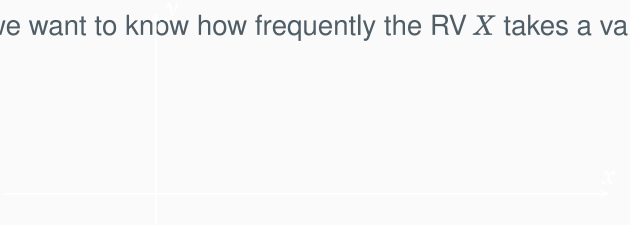
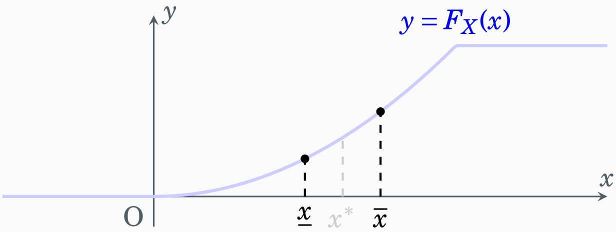


Figure:

CDF to the density.



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

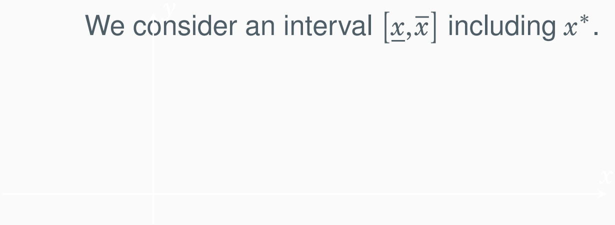
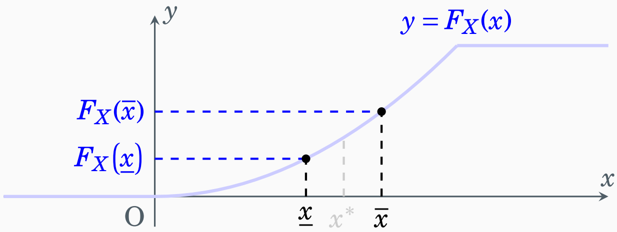


Figure:

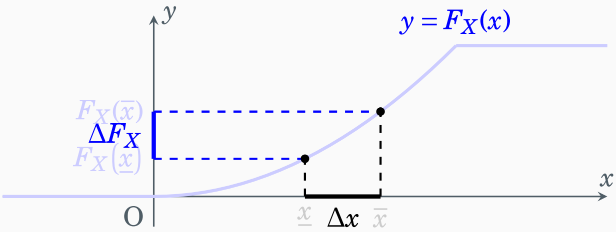
CDF to the density.



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

Figure:

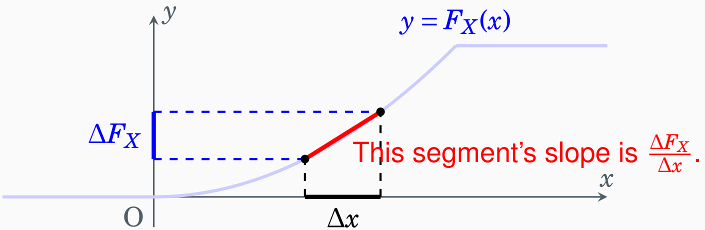
CDF to the density.



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

Figure:

CDF to the density.



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

The RV X tends to take a value around x^*

if the probability per length $\frac{\Delta F_X}{\Delta x}$, or the "density" is large.

CDF to the density.

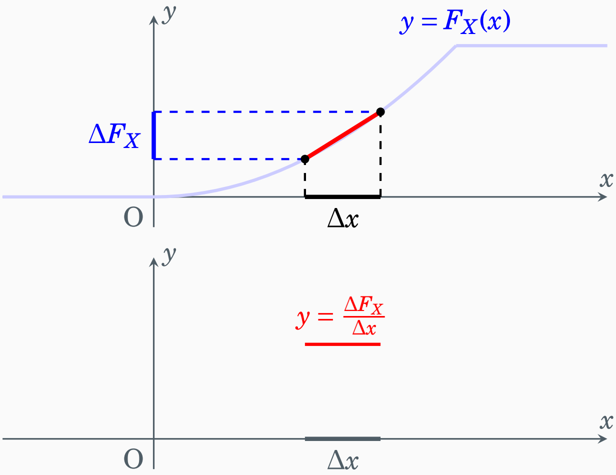


Figure: The graph plot of the probability per length $\frac{\Delta F_X}{\Delta x}$, or the “density”

CDF to the density.

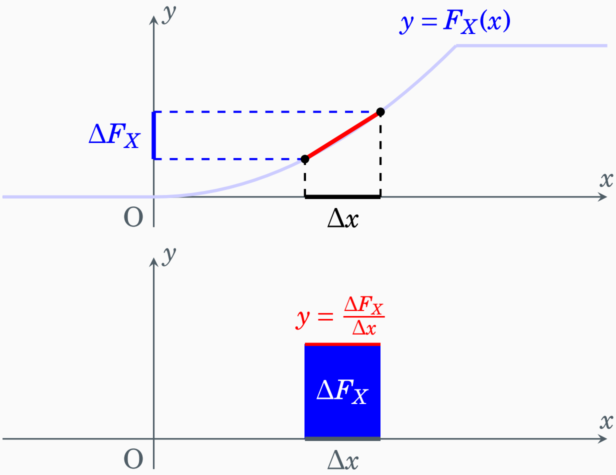


Figure: The area under the graph of $\frac{\Delta F_X}{\Delta x}$ is given by $\Delta x \cdot \frac{\Delta F_X}{\Delta x} = \Delta F_X$.

CDF to the density.

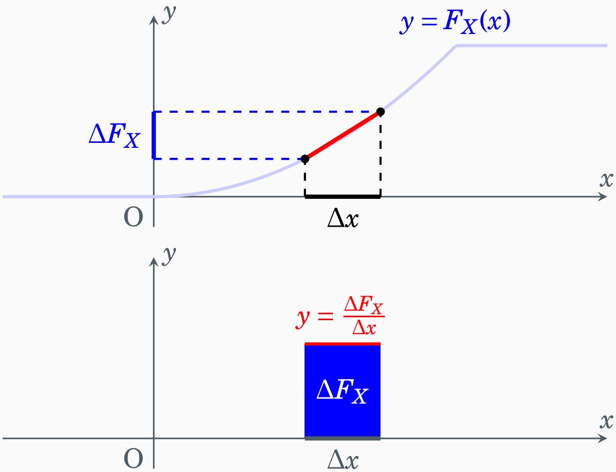


Figure: Let’s see how the density varies in other sections.

CDF to the density.

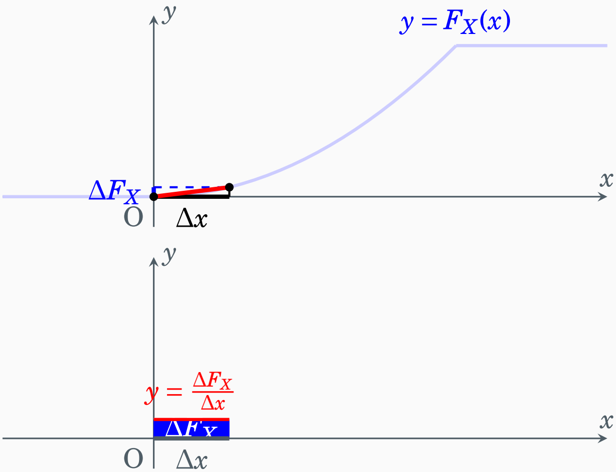


Figure: The density at the section $[0.00, 0.50]$ is small, reflecting the gentle slope of the graph of

© SUZUKI, Atsushi

CDF to the density.

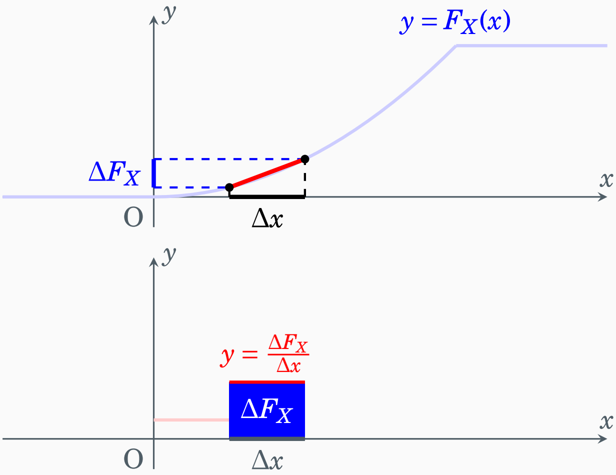


Figure: The density at the section $[0.50, 1.00]$ is larger, reflecting the steeper slope of the graph of

CDF to the density.

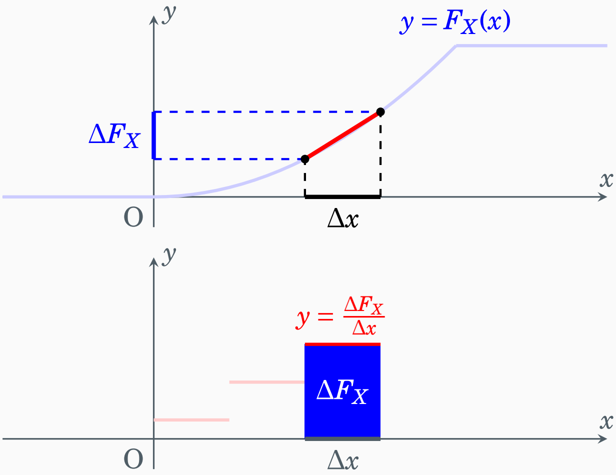


Figure: The density at the section $[1.00, 1.50]$ is even larger, reflecting the steeper slope of the

© SUZUKI, Atsushi

CDF to the density.

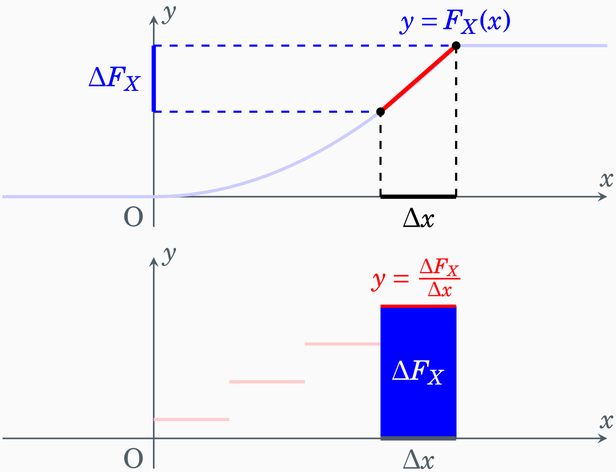


Figure: The density at the section $[1.50, 2.00]$ is even larger, reflecting the steeper slope of the

© SUZUKI, Atsushi

CDF to the density.

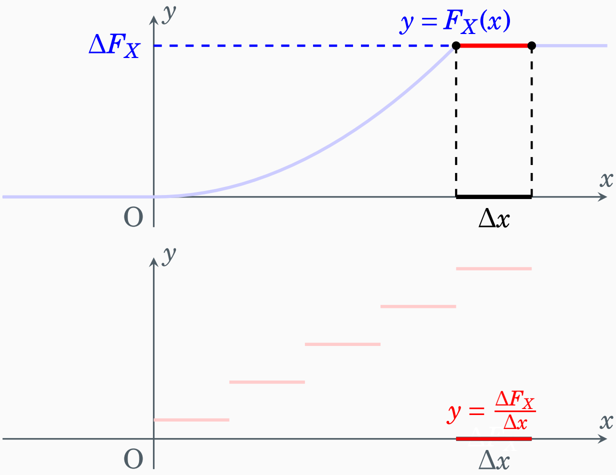


Figure: The density at the section $[2.00, 2.50]$ is zero since the graph of F_X is horizontal.

© SUZUKI, Atsushi

CDF to the density.

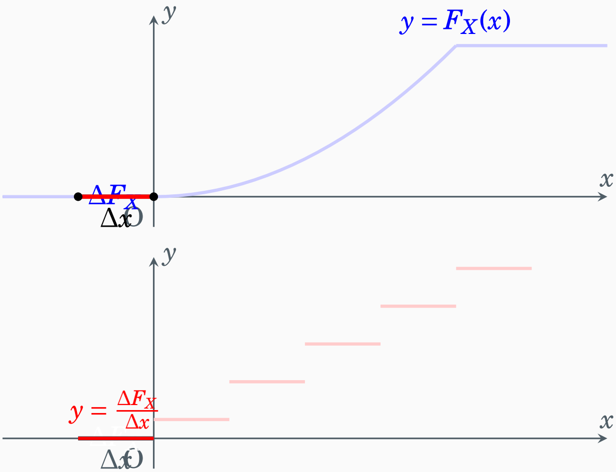


Figure: The density at the section $[-0.50, 0.90]$ is zero since the graph of F_X is horizontal.

CDF to the density.

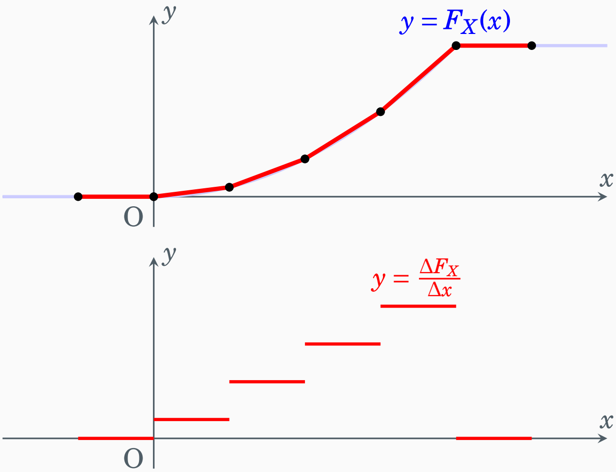


Figure: We have the graph of the piece-wise density of the distribution at each section.

CDF to the density.

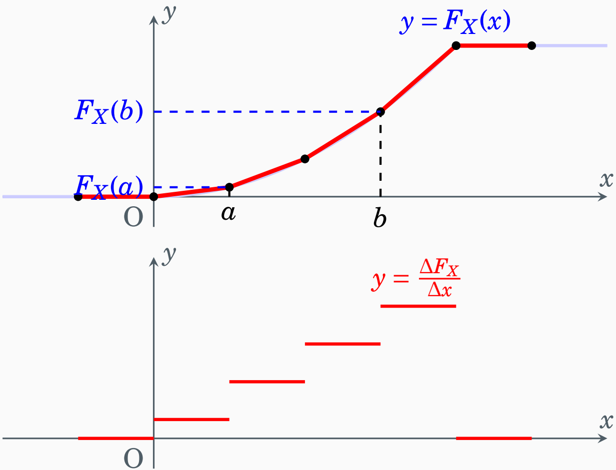


Figure: Using the piece-wise density, we can calculate $\Pr(a < X \leq b)$ if each of a and b is an end

CDF to the density.

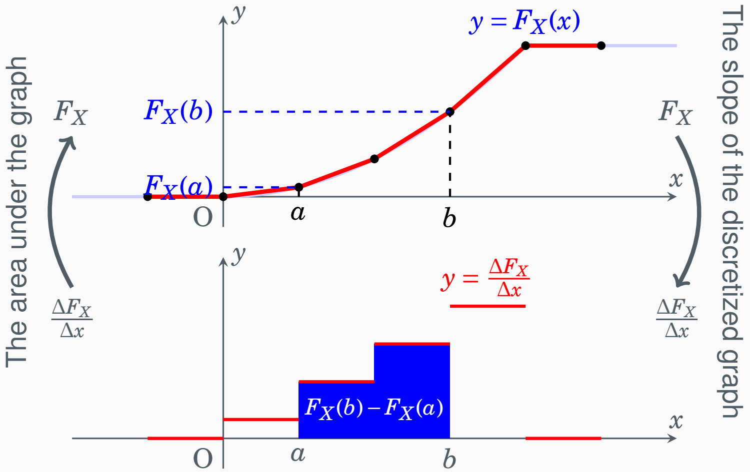


Figure: Actually, the probability $\Pr(a < X \leq b) = F_X(b) - F_X(a)$ is the area under the piece-wise

CDF to the density.

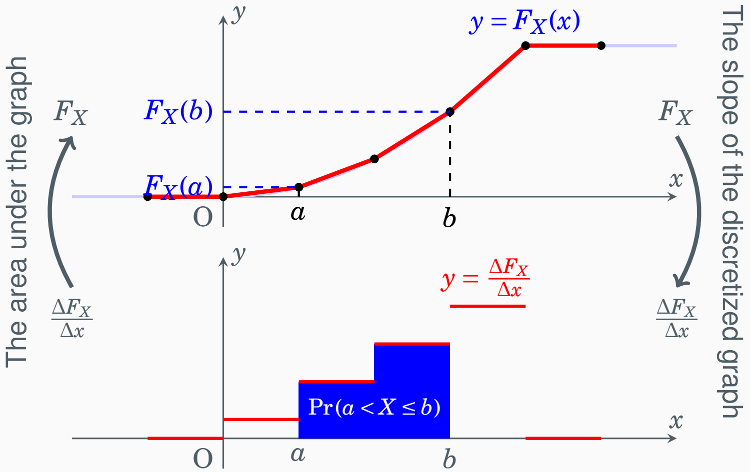


Figure: Actually, the probability $\Pr(a < X \le b) = F_X(b) - F_X(a)$ is the area under the piece-wise

CDF to the density.

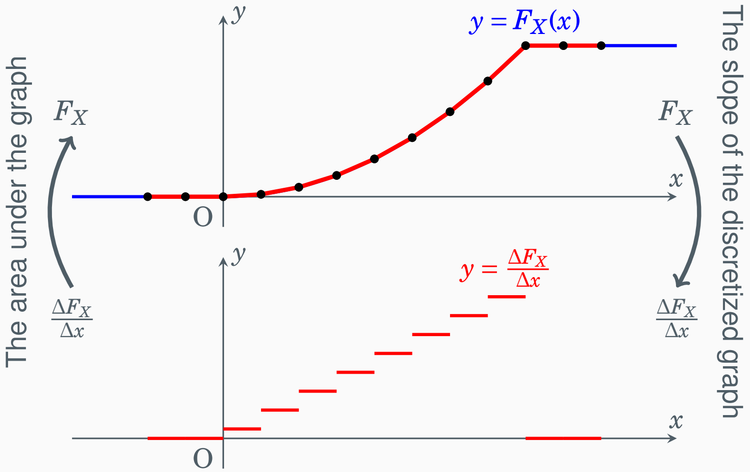


Figure: If the sections are shorter, we can evaluate $\Pr(a < X \leq b)$ for more pairs of a and b.

© SUZUKI, Atsushi

CDF to the density.

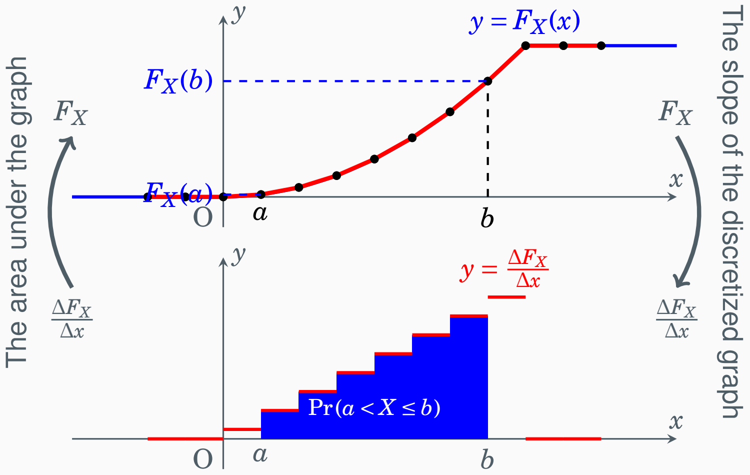


Figure: If the sections are shorter, we can evaluate $\Pr(a < X \leq b)$ for more pairs of a and b.

© SUZUKI, Atsushi

CDF to the density.

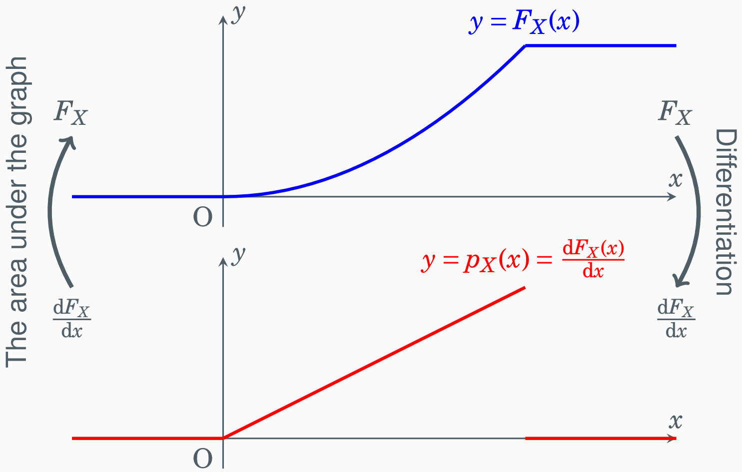


Figure: If we consider infinitely short sections, the density $\frac{\Delta F_X}{\Delta x}$ converges to the derivative $\frac{\mathrm{d}F_X}{\mathrm{d}x}$.

CDF to the density.

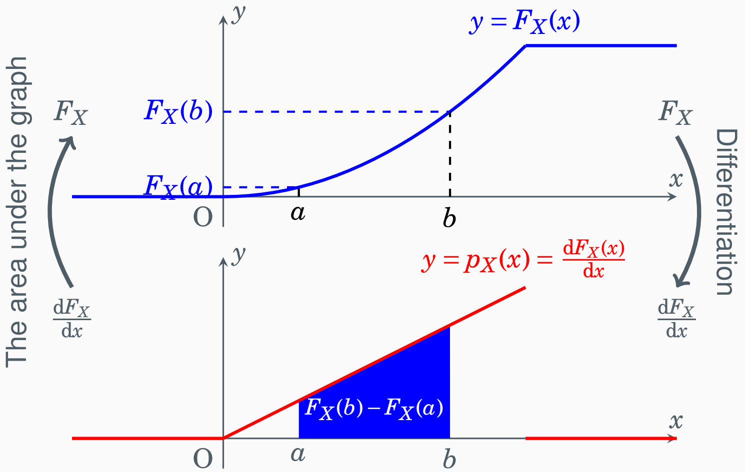


Figure: Actually, the probability $\Pr(a < X \leq b) = F_X(b) - F_X(a)$ is the area under the graph of $\frac{dF_X}{dx}$.

CDF to the density.

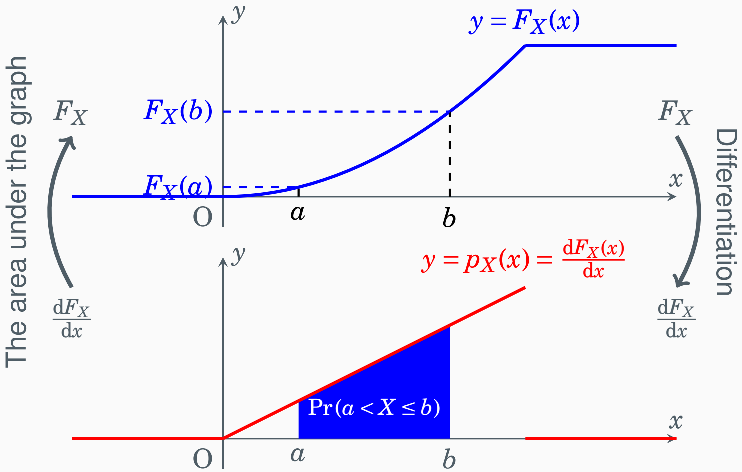


Figure: Actually, the probability $\Pr(a < X \le b) = F_X(b) - F_X(a)$ is the area under the graph of $\frac{dF_X}{dx}$.

Probability density function (PDF)

Definition (Probability density function and continuous random variable)

Let X be a RV. A function $p_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is called a ***probability density function (PDF)*** of X if the probability $\Pr(a < X \leq b)$ equals to the area bounded by the graph of $y = p_X(x)$ and $y = 0$ between $x = a$ and $x = b$ for all a and b such that $a \leq b$.
If a RV has at least one PDF, the RV is called a ***continuous random variable***.

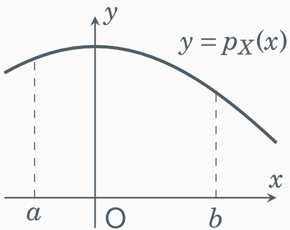


Figure: If p_X is a PDF of X, the probability $\Pr(a < X \leq b)$ is given by the area under the PDF in the domain $(a, b]$.

© SUZUKI, Atsushi

Probability density function (PDF)

Definition (Probability density function and continuous random variable)

Let X be a RV. A function $p_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is called a ***probability density function (PDF)*** of X if the probability $\Pr(a < X \leq b)$ equals to the area bounded by the graph of $y = p_X(x)$ and $y = 0$ between $x = a$ and $x = b$ for all a and b such that $a \leq b$.
If a RV has at least one PDF, the RV is called a ***continuous random variable***.

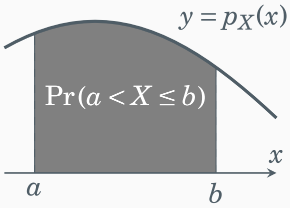


Figure: If p_X is a PDF of X, the probability $\Pr(a < X \leq b)$ is given by the area under the PDF in the domain $(a, b]$.

© SUZUKI, Atsushi

A continuous RV has nowhere a “mass.”

The area under a curve in a zero-length section is zero. Hence, if a RV is continuous, it has no probability mass anywhere. That is,

Theorem

If X is a continuous RV, the probability $\Pr(X = c)$ is zero for any $c \in \mathbb{R}$.

Hence, when we discuss a continuous RV, we do not need to discuss whether or not a section includes the endpoints. That is,

Corollary

Let X be a continuous RV and a and b be real values such that $a < b$. Then we have,

$$\Pr(a \leq X \leq b) = \Pr(a < X \leq b) = \Pr(a \leq X < b) = \Pr(a < X < b). \tag{38}$$

Hence, we can replace $a < X \leq b$ with $a \leq X \leq b$ or another in the definition of the PDF[6].

Note: the end-points are not ignorable for a discrete RV.

A discrete RV has a probability mass on any value in its support. Hence, for example, $\Pr(a \le X \le b) \ne \Pr(a < X \le b)$ in general.

For example, if X is the value when we roll an ideal six-sided dice, $\Pr(3 \le X \le 6) = \frac{4}{6} \ne \Pr(3 < X \le 6) = \frac{3}{6}$.

CDF and PDF

Assume that the CDF is differentiable at all the points on the real number line expect for finite points. As we can see in the construction of the PDF from the CDF, we can get the PDF by differentiating the CDF.

In practice, we usually know the PDF in advance but the CDF is unknown. Hence, we need to understand how to evaluate the area bounded by the graph of a general PDF.

© SUZUKI, Atsushi

Outline

3 **Continuous Random Variables**

Area, integration, and properties of PDF.

How to mathematically calculate the area under the curve?

Let X be a continuous RV and p_X be its PDF. Recall that the probability $\Pr(a \leq X \leq b)$ is given by the area under the graph of PDF p_X in the section $[a, b]$.

Hence, we need a mathematical tool to evaluate the area under the curve of a function in general.

Integration is the area to discuss the area under the graph of a function, (or the volume under the graph of a function in higher-dimensional space). We will learn it in the following.

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

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

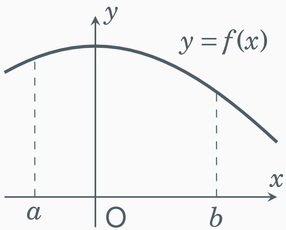


Figure: The definite integral is the area bounded by the graph of the function.

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

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

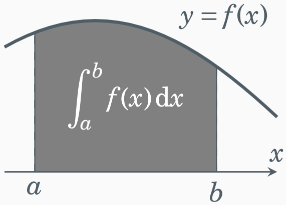


Figure: The definite integral is the area bounded by the graph of the function.

© SUZUKI, Atsushi

Definite Integral: When the function takes negative values

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

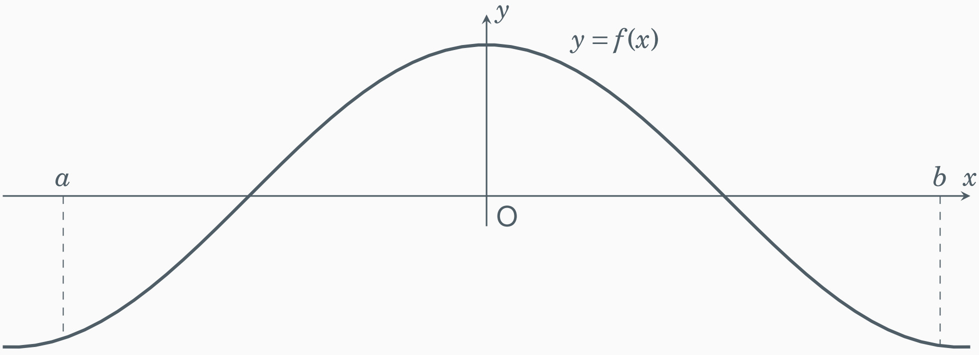


Figure: Areas bounded by the graph taking negative values are counted as negative values.

© SUZUKI, Atsushi

Definite Integral: When the function takes negative values

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

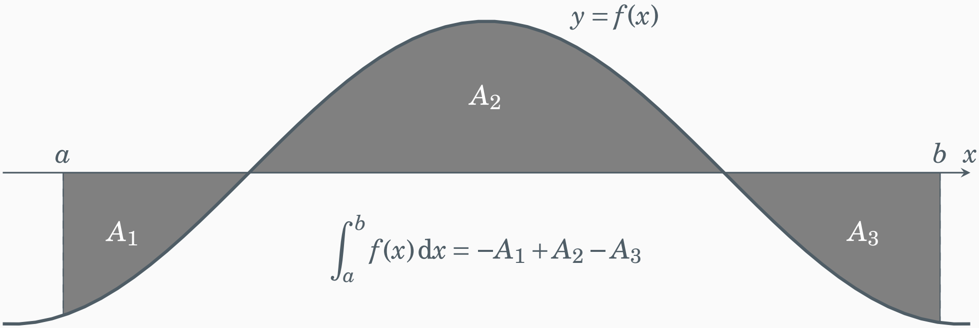


Figure: Areas bounded by the graph taking negative values are counted as negative values.

Any continuous RV has a nonnegative PDF

Assume that X is a continuous RV let p_X be a PDF of X. The probability $\Pr(a < X \leq b) = \int_a^b p_X(x)\,\mathrm{d}x$ is always nonnegative, so we expect the PDF p_X to be a nonnegative function.

Strictly speaking, a PDF of a RV is not unique, since the area bounded by the graph does not change even if we change the value of the function at finite or countable points[7].

Nevertheless, if a RV has a PDF, we can assume that it is a nonnegative function without loss of generality.

Theorem

*Let X be a continuous RV, i.e., there is a PDF of X. Then, there exists a **nonegative** PDF of X, i.e, a PDF p_X such that $p_X(x) \geq 0$ at any $x \in \mathbb{R}$.*

In the following, we always assume that a PDF is nonnegative.

[7] Strictly speaking, at points in a set with measurement zero.

© SUZUKI, Atsushi

Probability of a RV being in a complicated shape

If we want to calculate the probability $\Pr(X \in A)$, where $A \subset \mathbb{R}$ has a complicated shape, we can calculate it using the sum rule.

Specifically, suppose that we have a decomposition $A = \bigcup_{i=1}^{n} (a_i, b_i]$, where $(a_i, b_i] \cap (a_j, b_j] = \emptyset$. Then, we have that

$$\Pr(X \in A) = \sum_{i=1}^{n} \Pr(a_i < X \leq b_i). \tag{39}$$

If X is a continuous RV and p_X is its PDF, the above value equals $\sum_{i=1}^{n} \int_{a_i}^{b_i} p_X(x) \mathrm{d}x$.

The same discussion holds even if the decomposition includes open sections like (a_i, b_i) or closed sections like $[a_j, b_j]$.

Note that the above calculation is not always correct if the decomposition includes an uncountably infinite number of sections.

Probability of a RV being in an infinite length section

If we need to evaluate the probability $\Pr(a < X)$, what we do is consider $\Pr(a < X \le b)$ for an infinitely large b. Hence, we have that $\Pr(a < X) = \lim_{b \to +\infty} \Pr(a < X \le b)$. The reverse holds for $\Pr(X \le b)$. In other words, we can evaluate those probabilities by taking the limit of a definite integral as follows.

Theorem

Let X be a continuous RV, whose PDF is p_X, and a and b be real values. Then,

- $\Pr(a < X) = \Pr(a \le X) = \lim_{b \to +\infty} \int_a^b p_X(x) \mathrm{d}x,$
- $\Pr(X < b) = \Pr(X \le b) = \lim_{a \to -\infty} \int_a^b p_X(x) \mathrm{d}x.$

© SUZUKI, Atsushi

The "sum" of the PDF is one.

The section $(a, 0]$ includes all the nonpositive numbers if a is infinitely small and the section $(0, b]$ includes all the positive numbers b is infinitely large. Since a continuous RV X always takes a real value, the sum of the probabilities $\Pr(a < X \leq 0) + \Pr(0 < X \leq b)$ is 1 if a is infinitely small and b is infinitely large. Hence, the following always hold.

Theorem

Let X be a continuous RV whose PDF is p_X. We have that

$$\lim_{a \to -\infty} \int_a^0 p_X(x)\,\mathrm{d}x + \lim_{b \to +\infty} \int_0^b p_X(x)\,\mathrm{d}x = 1 \tag{40}$$

The above property is similar to a property of the probability mass function (PMF) of a discrete RV. To see that, we will introduce the ***improper integral***.

Improper integral

As we have seen, we often want to calculate limits of the definite integral. We call them ***improper integrals***, and use special notations as follows.

Definition (Improper integrals)

Let $f : \mathbb{R} \to \mathbb{R}$ be a function and a and b be real values. We define the value $\int_a^{+\infty} f(x)\,\mathrm{d}x$, $\int_{-\infty}^{b} f(x)\,\mathrm{d}x$, and $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x$ by the following.

- $\displaystyle\int_a^{+\infty} f(x)\,\mathrm{d}x := \lim_{b\to+\infty} \int_a^{b} f(x)\,\mathrm{d}x,$
- $\displaystyle\int_{-\infty}^{b} f(x)\,\mathrm{d}x := \lim_{a\to-\infty} \int_a^{b} f(x)\,\mathrm{d}x,$
- $\displaystyle\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x := \int_{-\infty}^{0} f(x)\,\mathrm{d}x + \int_0^{+\infty} f(x)\,\mathrm{d}x.$

Interpretation of improper integrals

We can regard improper integrals $\int_a^{+\infty} f(x)\,\mathrm{d}x$, $\int_{-\infty}^{b} f(x)\,\mathrm{d}x$, and $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x$ as the signed areas bounded by a graph of f in the section $(a, +\infty)$, $(-\infty, b)$, and $(-\infty, +\infty)$, respectively.

Rewriting properties of the PDF using improper integrals

We can rewrite the properties of the PDF in previous slides as follows.

Theorem

Let X be a continuous RV, whose PDF is p_X, and a and b be real values. Then,

- $\Pr(a < X) = \Pr(a \le X) = \int_{a}^{+\infty} p_X(x)\,\mathrm{d}x,$
- $\Pr(X < b) = \Pr(X \le b) = \int_{-\infty}^{b} p_X(x)\,\mathrm{d}x,$
- $\int_{-\infty}^{+\infty} p_X(x)\,\mathrm{d}x = 1.$

The third property is similar to a property of the PMF: $\sum_{x \in \mathscr{X}} P_X(x) = 1$, where X is a discrete RV, $\mathscr{X}$ is its support and P_X is its PMF.

Properties of the definite integral

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

- $\displaystyle \int_b^a f(x)\,\mathrm{d}x := -\int_a^b f(x)\,\mathrm{d}x \qquad \text{(by definition)},$
- $\displaystyle \int_a^a f(x)\,\mathrm{d}x = 0 \qquad \text{(The area is zero in a zero length section.)}.$
- $\displaystyle \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x \qquad \text{(horizontal concatenation)}.$

Outline

3 **Continuous Random Variables**

Calculating integral

Calculating definite integrals

Let X be a continuous RV and p_X be its PDF. Since the probability $\Pr(a < X \leq b)$ is given by the definite integral $\int_a^b p(x)\,\mathrm{d}x$, we need to know **how to calculate definite integrals** to understand the behavior of the continuous RV X.

[8] e.g., the trapezoidal rule, the Gauss-Legendre quadrature rule, the double exponential formula

Calculating definite integrals

Let X be a continuous RV and p_X be its PDF. Since the probability $\Pr(a < X \leq b)$ is given by the definite integral $\int_a^b p(x)\,\mathrm{d}x$, we need to know **how to calculate definite integrals** to understand the behavior of the continuous RV X.

There are two directions to calculate definite integrals.

- **Numerical integration** by approximating the area by shapes of which we can calculate the area easier.
- **Analytical integration** by conducting integration as the inverse operation of differentiation.

[8] e.g., the trapezoidal rule, the Gauss-Legendre quadrature rule, the double exponential formula

Calculating definite integrals

Let X be a continuous RV and p_X be its PDF. Since the probability $\Pr(a < X \leq b)$ is given by the definite integral $\int_a^b p(x)\,\mathrm{d}x$, we need to know **how to calculate definite integrals** to understand the behavior of the continuous RV X.

There are two directions to calculate definite integrals.

- **Numerical integration** by approximating the area by shapes of which we can calculate the area easier.
- **Analytical integration** by conducting integration as the inverse operation of differentiation.

In general, numerical integration methods[8] can apply to a variety of cases but cause an approximation error. The analytical integration methods can give us the exact value but have limited applications. In practice, we combine them depending on the situation. In this lecture, **we focus on analytical methods**. It also helps us learn numerical integration.

[8] e.g., the trapezoidal rule, the Gauss-Legendre quadrature rule, the double exponential formula

Basic idea of calculating an integral

When we constructed the probability density function (PDF) p_X, we differentiated the cumulative distribution function (CDF) F_X. Specifically, $p_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$. Conversely, we observed that the area $\int_a^b p_X(x)\,\mathrm{d}x$ under the graph of the PDF corresponds to the difference $F_X(b) - F_X(a)$.

Basic idea of calculating an integral

When we constructed the probability density function (PDF) p_X, we differentiated the cumulative distribution function (CDF) F_X. Specifically, $p_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$. Conversely, we observed that the area $\int_a^b p_X(x)\,\mathrm{d}x$ under the graph of the PDF corresponds to the difference $F_X(b) - F_X(a)$.

To wrap up, to calculate the definite integral $\int_a^b p_X(x)\,\mathrm{d}x$, we can use a function whose derivative is p_X.

Basic idea of calculating an integral

When we constructed the probability density function (PDF) p_X, we differentiated the cumulative distribution function (CDF) F_X. Specifically, $p_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$. Conversely, we observed that the area $\int_a^b p_X(x)\,\mathrm{d}x$ under the graph of the PDF corresponds to the difference $F_X(b) - F_X(a)$.

To wrap up, to calculate the definite integral $\int_a^b p_X(x)\,\mathrm{d}x$, we can use a function whose derivative is p_X.

According to the ***fundamental theorem of calculus (FTC)***, this relation between the derivative and the definite integral applies to a general function. We can use this relation to calculate a definite integral.

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 primitive function $F : [a,b] \to \mathbb{R}$ of f, then we have that

$$\int_a^b f(t)\,\mathrm{d}t = F(b) - F(a). \tag{41}$$

We often denote $F(b) - F(a)$ by $[F(x)]_a^b$.

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

Calculating the definite integral

To calculate the definite integral

$$\int_a^b f(x)\,\mathrm{d}x\,, \tag{42}$$

the following steps suffice.

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

© SUZUKI, Atsushi

Examples of calculating a definite integral

Example

Let $f(x) = x$.

We can calculate the definite integral $\displaystyle\int_{-4}^{5} f(x)\,\mathrm{d}x = \int_{-4}^{5} x\,\mathrm{d}x$ as follows.

- **Step 1**:

- **Step 2**:

Examples of calculating a definite integral

Example

Let $f(x) = x$.

We can calculate the definite integral $\int_{-4}^{5} f(x)\,\mathrm{d}x = \int_{-4}^{5} x\,\mathrm{d}x$ as follows.

- **Step 1**: Find a primitive (antiderivative) function F, which satisfies $F' = f$. In this example case, we can use a function $F(x) = \frac{1}{2}x^2$ as a primitive function since $\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{2}x^2 = x$.
- **Step 2**:

Examples of calculating a definite integral

Example

Let $f(x) = x$.

We can calculate the definite integral $\int_{-4}^{5} f(x)\,\mathrm{d}x = \int_{-4}^{5} x\,\mathrm{d}x$ as follows.

- **Step 1**: Find a primitive (antiderivative) function F, which satisfies $F' = f$. In this example case, we can use a function $F(x) = \frac{1}{2}x^2$ as a primitive function since $\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{2}x^2 = x$.
- **Step 2**: Evaluate the value of $[F(x)]_{-4}^{5} := F(5) - F(-4)$. In this example case, $F(5) - F(-4) = \frac{1}{2}(5)^2 - \frac{1}{2}(-4)^2 = \frac{25}{2} - 8 = \frac{9}{2}$.

Examples of calculating a definite integral

Example

Let $f(x) = x$.

We can calculate the definite integral $\int_{-4}^{5} f(x)\,\mathrm{d}x = \int_{-4}^{5} x\,\mathrm{d}x$ as follows.

- **Step 1**: Find a primitive (antiderivative) function F, which satisfies $F' = f$. In this example case, we can use a function $F(x) = \frac{1}{2}x^2$ as a primitive function since $\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{2}x^2 = x$.
- **Step 2**: Evaluate the value of $[F(x)]_{-4}^{5} := F(5) - F(-4)$. In this example case, $F(5) - F(-4) = \frac{1}{2}(5)^2 - \frac{1}{2}(-4)^2 = \frac{25}{2} - 8 = \frac{9}{2}$.

Hence, we have that

$$\int_{-4}^{5} f(x)\,\mathrm{d}x = \frac{9}{2}. \tag{43}$$

A primitive function is not unique.

As we have seen, finding a primitive function is essential to calculate the definite integral. Here, we must note that 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 $f(x) = x$.

The primitive function is unique up to an additive constant.

A primitive function is not unique. **However**, it is unique **up to an additive constant** in the following sense.

The primitive function is unique up to an additive constant.

A primitive function is not unique. **However**, it is unique **up to an additive constant** in the following sense.

Theorem (The primitive function is unique up to an additive constant)

Let a *and* b *be real values such that* $a < b$*. If both* $F_1 : [a,b] \to \mathbb{R}$ *and* $F_2 : [a,b] \to \mathbb{R}$ *are primitive functions of* f*, the difference between* F_1 *and* F_2 *is a constant function. In other words, there exists a constant* $C \in \mathbb{R}$ *such that* $F_2(x) - F_1(x) = C$*.*

The primitive function is unique up to an additive constant.

A primitive function is not unique. **However**, it is unique **up to an additive constant** in the following sense.

Theorem (The primitive function is unique up to an additive constant)

Let a and b be real values such that $a < b$. If both $F_1 : [a,b] \to \mathbb{R}$ and $F_2 : [a,b] \to \mathbb{R}$ are primitive functions of f, the difference between F_1 and F_2 is a constant function. In other words, there exists a constant $C \in \mathbb{R}$ such that $F_2(x) - F_1(x) = C$.

To wrap up, if F is a primitive function of f, then, for any constant C, the function given by $F(x) + C$ is also a primitive function of f, and conversely, all the primitive functions are written in this form. We write this fact as follows.

$$\int f(x)\,\mathrm{d}x = F(x) + C, \tag{44}$$

Here, the symbol $\int f(x)\,\mathrm{d}x$ in the LHS denotes all the primitive functions of f. Here, the constant C in the RHS is called the ***constant of integration***.

© SUZUKI, Atsushi

Examples of primitive functions

Example

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

Note about the proof

We can prove the uniqueness of the primitive function up to an additive constant by the ***mean value theorem***.

Indefinite integral

Let f be a function and a be a real value. The function defined by the following form is called an ***indefinite integral*** of f.

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

It is known that if f be continuous, then an indefinite integral is a primitive function of f. Note that some literature use the term "indefinite integral" to refer to a primitive function for this reason, while not all primitive functions are written in the above form.

© SUZUKI, Atsushi

Linearity of the antidifferentiation and integral

Since the derivation is a linear operator, the antidifferentiation, the operation to find a primitive function, is linear as well in the following sense.

Theorem (Linearity of antidifferentiation)

Let $f, g : \mathbb{R} \to \mathbb{R}$ be functions and F and G be the primitive functions of f and g, respectively. Also, let α and β be real values.
Then, $\alpha F + \beta G$ is a primitive function of $\alpha f + \beta g$.
In other words,

$$\int \left(\alpha f(x) + \beta g(x)\right) \mathrm{d}x = \alpha \int f(x) \, \mathrm{d}x + \beta \int g(x) \, \mathrm{d}x \,. \tag{46}$$

We can easily prove the above by taking the derivatives of both sides[9].

[9] Strictly speaking, we should consider the uniqueness of the primitive function up to an additive constant

Linearity of the definite integral

By combining the linearity of the antidifferentiation and the FTC, we can immediately get the linearity of the definite integral, which is a useful formula.

Corollary (Linearity of the definite integral)

Let $f, g : \mathbb{R} \to \mathbb{R}$ *be functions and* a, b, α *and* β *be real values. Then,*

$$\int_a^b \big(\alpha f(x) + \beta g(x)\big)\,\mathrm{d}x = \alpha \int_a^b f(x)\,\mathrm{d}x + \beta \int_a^b g(x)\,\mathrm{d}x\,. \tag{47}$$

Example of the linearity of antidifferentiation

Example

In the following, C is the constant of integration.

- $\displaystyle\int (\cos x + x^2)\,\mathrm{d}x = \int \cos x\,\mathrm{d}x + \int x^2\,\mathrm{d}x = \sin x + \frac{1}{3}x^3 + C$. Hence, $\displaystyle\int_0^\pi (\cos x + x^2)\,\mathrm{d}x = \left(\sin\pi + \frac{1}{3}\pi^3\right) - \left(\sin 0 + \frac{1}{3}\cdot 0^3\right) = \frac{1}{3}\pi^3$.
- $\displaystyle\int 5\exp(x)\,\mathrm{d}x = 5\int \exp(x)\,\mathrm{d}x = 5\exp(x) + C$. Hence, $\displaystyle\int_1^3 5\exp(x)\,\mathrm{d}x = (5\exp(3)) - (5\exp(1)) = 5\mathrm{e}(\mathrm{e}^2 - 1)$.

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, then, e.g.,

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

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

However, generally speaking, antidifferentiation is more difficult than differentiation. Specifically, we have no formulae to find a primitive function of a general product or composition like in differentiation. Nevertheless, we have some techniques to make such calculation more feasible for some cases, called ***integration by parts*** and ***integration by substitution***.

© SUZUKI, Atsushi

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

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

© SUZUKI, Atsushi

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.

© SUZUKI, Atsushi

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.

© SUZUKI, Atsushi

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)

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

Example of definite integral calculation in probability theory

Example (Exponential distribution)

The distribution of a RV X is called the ***exponential distribution*** with mean μ if it has a PDF p_X given by

$$p_X(x) := \begin{cases} 0 & \text{if } x < 0, \\ \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right) & \text{if } x \geq 0. \end{cases} \tag{54}$$

For nonnegative numbers a and b, the probability $\Pr(a < X \leq b)$ is given by

$$\begin{aligned} \Pr(a < X \leq b) = \int_a^b p_X(x)\,\mathrm{d}x &= \int_a^b \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right)\mathrm{d}x \\ &= \left[-\exp\left(-\frac{x}{\mu}\right)\right]_a^b = \exp\left(-\frac{a}{\mu}\right) - \exp\left(-\frac{b}{\mu}\right). \end{aligned} \tag{55}$$

A primitive function of the product/composition is not easily found.

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

$$\int \frac{1}{x}\,\mathrm{d}x = \log|x| + C, \int \sin x\,\mathrm{d}x = -\cos + C, \int \left(-x^2\right)\mathrm{d}x = -\frac{1}{3}x^3 + C, \int \exp(x)\,\mathrm{d}x = \exp(x) + C. \tag{56}$$

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 $\log x$.

A primitive function of the product/composition is not easily found.

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

$$\int \frac{1}{x}\,\mathrm{d}x = \log|x| + C, \int \sin x\,\mathrm{d}x = -\cos + C, \int \left(-x^2\right)\mathrm{d}x = -\frac{1}{3}x^3 + C, \int \exp(x)\,\mathrm{d}x = \exp(x) + C. \tag{56}$$

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 implemented, you might need to calculate the definite integral using a numerical method.

A primitive function of the product/composition is not easily found.

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

$$\int \frac{1}{x}\,\mathrm{d}x = \log|x| + C, \int \sin x\,\mathrm{d}x = -\cos + C, \int \left(-x^2\right)\mathrm{d}x = -\frac{1}{3}x^3 + C, \int \exp(x)\,\mathrm{d}x = \exp(x) + C. \tag{56}$$

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 $\exp(-x^2)$. 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

3 **Continuous Random Variables**

- Summary statistics of continuous RV and integral

Expectation (mean) of a continuous random variable

The expectation of a continuous RV is defined similarly to that of a discrete RV. Specifically, we get the definition for a continuous RV by replacing the PMF and the sum with the PDF and the integration in the definition for a discrete RV.

Expectation (mean) of a continuous random variable

The expectation of a continuous RV is defined similarly to that of a discrete RV. Specifically, we get the definition for a continuous RV by replacing the PMF and the sum with the PDF and the integration in the definition for a discrete RV.

Definition (Expectation of a continuous RV)

Let X be a continuous RV and p_X be its probability density function (PDF). Then, the expectation $\mathbb{E}X$ of X is defined by

$$\mathbb{E}X := \int_{-\infty}^{+\infty} x p(x) \, \mathrm{d}x \,. \tag{57}$$

Cf.) The expectation of a discrete RV X is given by $\sum_{x \in \mathcal{X}} x P_X(x)$, where P_X is the probability mass function.

Example: expectation of exponential distribution

Example (Expectation of the exponential distribution)

The PDF p_X of a RV X following the exponential distribution with mean μ is given by

$$p_X(x) := \begin{cases} 0 & \text{if } x < 0, \\ \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right) & \text{if } x \geq 0. \end{cases} \tag{58}$$

Noting that the density is zero for the negative domain, we can calculate the expectation $\mathbb{E}X$ using integration by parts as follows.

$$\begin{aligned} \mathbb{E}X &= \int_{-\infty}^{+\infty} x p_X(x)\,\mathrm{d}x = \int_0^{+\infty} x \cdot \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right)\mathrm{d}x = \int_0^{+\infty} x\left(-\exp\left(-\frac{x}{\mu}\right)\right)'\mathrm{d}x \\ &= \left[x \cdot \left(-\exp\left(-\frac{x}{\mu}\right)\right)\right]_0^{+\infty} - \int_0^{+\infty} (x)' \cdot \left(-\exp\left(-\frac{x}{\mu}\right)\right)\mathrm{d}x = -\int_0^{+\infty}\left(-\exp\left(-\frac{x}{\mu}\right)\right)\mathrm{d}x \\ &= -\left[\mu\exp\left(-\frac{x}{\mu}\right)\right]_0^{+\infty} = \mu. \end{aligned} \tag{59}$$

The expectation of a function of a continuous RV

A function of a discrete RV is always a discrete RV. However, a function of a continuous RV is not always a continuous RV.

The expectation of a function of a continuous RV

A function of a discrete RV is always a discrete RV. However, a function of a continuous RV is not always a continuous RV.

For example, if f is the sign function defined by

$$f(x) := \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ +1 & \text{if } x > 0, \end{cases} \tag{60}$$

and X is a continuous RV whose PDF p_X is given by

$$p_X(x) = \begin{cases} +1 & \text{if } -\frac{1}{2} \leq x \leq +\frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{61}$$

Then, the RV $f(X)$ takes values -1 and $+1$ with equal probability. In particular, it is a discrete RV, whose support is $\{-1, +1\}$.

© SUZUKI, Atsushi

The expectation of a function of a continuous RV

A function of a discrete RV is always a discrete RV. However, a function of a continuous RV is not always a continuous RV.

Even though a function of a continuous RV may not be a continuous RV, its expectation can always be calculated by the following formula, which is similar to the formula for a discrete RV.

Theorem

Let X be a continuous RV and its PDF be p_X. Also, let $f : \mathbb{R} \to \mathbb{R}$ be a real-valued function taking a real value as an input. The expectation $\mathbb{E} f(X)$ of the random variable X is given as follows.

$$\mathbb{E} f(X) = \int_{-\infty}^{+\infty} f(x) p_X(x) \, \mathrm{d}x \,. \tag{60}$$

Cf.) For a discrete RV whose support and PMF are $\mathscr{X}$ and P_X, respectively, we have that $\mathbb{E} f(X) = \sum_{x \in \mathscr{X}} f(x) P_X(x)$.

© SUZUKI, Atsushi

The linearity of the expectation on continuous RVs

The following theorem, which holds for a discrete RV, also holds for a continuous RV.

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

$$\mathbb{E}[af(X) + bg(X)] = a\,\mathbb{E}f(X) + b\,\mathbb{E}g(X). \tag{61}$$

Variance and standard deviation of a continuous random variable

The definitions of the variance and standard deviation are the same for a continuous RV. Specifically, for a continuous RV X, whose expectation is μ_X, its variance $\mathbb{V}(X)$ is defined by $\mathbb{V}(X) := \mathbb{E}\left(X - \mu_X\right)^2$. The standard deviation is defined by $\sigma_X := \sqrt{\mathbb{V}(X)}$.

When we know the explicit form of the PDF, we can use the following formulae.

Theorem

Let X be a continuous RV and its PDF be p_X. Suppose that the expectation $\mathbb{E}X = \int_{-\infty}^{+\infty} x p_X(x) \, \mathrm{d}x$ exists and denote it by μ_X. The variance $\mathbb{V}(X)$ is given by the following formula.

$$\mathbb{V}(X) = \int_{-\infty}^{+\infty} \left(x - \mu_X\right)^2 p_X(x) \, \mathrm{d}x = \int_{-\infty}^{+\infty} x^2 p_X(x) \, \mathrm{d}x - \left(\mu_X\right)^2. \tag{62}$$

Outline

3 Continuous Random Variables

- Jointly continuous random variables and multiple integral

Handling multiple non-discrete random variables

Similar to the univariate random variable case, we can define the cumulative distribution function (CDF) for multiple RV even if they are not discrete.

Definition (The CDF of two RVs)

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

$$F_{X,Y}(x,y) := \Pr(X \leq x \wedge Y \leq y), \tag{63}$$

where $\wedge$ indicates the logical "and" statement.

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

$$\Pr(a_1 < X \leq b_1 \wedge a_2 < Y \leq b_2) = F_{X,Y}(b_1, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(b_1, a_2) + F_{X,Y}(b_1, b_2) \tag{64}$$

To define the probability density function for a multivariate RV

Let $X_1, X_2, \ldots, X_m$ be random variables. As in the univariate random variable case, the CDF may not be easy to interpret or not be elementary even in practical cases. Hence, we want to define the probability density function (PDF) for multiple RV cases.

© SUZUKI, Atsushi

To define the probability density function for a multivariate RV

Let $X_1, X_2, \ldots, X_m$ be random variables. As in the univariate random variable case, the CDF may not be easy to interpret or not be elementary even in practical cases. Hence, we want to define the probability density function (PDF) for multiple RV cases.

The univariate continuous RV theory allows us to define the PDF for each RV, but they are not sufficient to understand the behavior of a multiple RVs completely, as we saw in multiple discrete RV cases. To tackle this issue, for discrete RV cases, we evaluated the Joint PMF, which returns the probability mass of the event $(X_1, X_2, \ldots, X_m) = (x_1, x_2, \ldots, x_m)$. Similarly, we want to define the function that returns the probability density at $(X_1, X_2, \ldots, X_m) = (x_1, x_2, \ldots, x_m)$. Since the PDF for a univariate continuous RV was defined using the area under the graph, let us define the graph of a multivariate function and the high-dimensional area (volume) in the following.

The graph of a multivariate function and 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

$$\{(x_1, x_2, \ldots, x_m, f(x_1, x_2, \ldots, x_m)) | (x_1, x_2, \ldots, x_m) \in D\} \tag{65}$$

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 $\int_D f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$.

© SUZUKI, Atsushi

Joint PDF

Based on the definition of multiple integration, we can define the joint probability density function (joint PDF) of a multivariate random variable.

Definition

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

$$\Pr((X_1, X_2, \ldots, X_m) \in D) = \int_D p_{X_1,X_2,\ldots,X_m}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{66}$$

for any m-dimensional hyper-rectangle D, then the function $p_{X_1,X_2,\ldots,X_m}$ is called the ***joint probability density function (joint PDF)*** of $X_1, X_2, \ldots, X_m$.
If $(X_1, X_2, \ldots, X_m)$ have a joint PDF, we call them ***jointly continuous random variables (jointly continuous RVs)***.

© SUZUKI, Atsushi

Multiple continuous RVs are not always jointly continuous

Let $X_1, X_2, \ldots, X_m$ be continuous RVs. In other words, suppose that there exist PDFs $p_{X_1}, p_{X_2}, \ldots, p_{X_m}$ for $X_1, X_2, \ldots, X_m$, respectively.

Even under this assumption, **it is possible that $X_1, X_2, \ldots, X_m$ have no joint PDF**.

© SUZUKI, Atsushi

Multiple continuous RVs are not always jointly continuous

Let $X_1, X_2, \ldots, X_m$ be continuous RVs. In other words, suppose that there exist PDFs $p_{X_1}, p_{X_2}, \ldots, p_{X_m}$ for $X_1, X_2, \ldots, X_m$, respectively.

Even under this assumption, **it is possible that $X_1, X_2, \ldots, X_m$ have no joint PDF**.

Example

For example, let X and Y be a continuous RV following the uniform distribution on $[0,1]$ and suppose that $X = Y$ always hold. Then, both X and Y have the same PDF $p_X(z) = p_Y(z) \begin{cases} 1 & \text{if } 0 \leq z \leq 1, \\ 0 & \text{otherwise.} \end{cases}$, so both X and Y are continuous RVs. However, the probability mass concentrates on the segment from the origin $(0,0)$ to the point $(1,1)$ in the xy space. The segment has zero area, but if there existed the joint PDF, the volume bounded by the graph of the joint PDF on the segment would be 1. This is a contradiction, so X, Y have no joint PDF. Hence, X, Y are not jointly continuous.

© SUZUKI, Atsushi

Multiple continuous RVs are not always jointly continuous

Let $X_1, X_2, \ldots, X_m$ be continuous RVs. In other words, suppose that there exist PDFs $p_{X_1}, p_{X_2}, \ldots, p_{X_m}$ for $X_1, X_2, \ldots, X_m$, respectively.

Even under this assumption, **it is possible that $X_1, X_2, \ldots, X_m$ have no joint PDF**.

To wrap up, **even if both X and Y are continuous RVs, it does not follow that the X, Y are jointly continuous!** Conversely, jointly continuous RVs are always multiple continuous RVs.

For this reason, we need to **distinguish multiple continuous RVs and jointly continuous RVs**. The former is the broader concept, but we focus on the latter since we have many mathematical tools based on the multiple integration to analyze them.

© SUZUKI, Atsushi

Multiple integral on a complicated shape

In high-dimensional space, we might want to consider the volume bounded by a function in a complicated shape, say $\mathscr{A}$, that cannot be represented as a union of hyper-rectangles.

Multiple integral on a complicated shape

In high-dimensional space, we might want to consider the volume bounded by a function in a complicated shape, say $\mathscr{A}$, that cannot be represented as a union of hyper-rectangles.

For example, we might want to consider the probability $\Pr((X,Y) \in \mathscr{A})$, where $\mathscr{A} := \left\{(x,y) \in \mathbb{R}^2 \middle| x^2 + y^2 \leq 1\right\}$ is defined as the unit disk centered at the origin. We cannot decompose the disk into rectangles, so we cannot evaluate the probability by the sum rule if we can only define the probability of the multivariate RV being in a rectangle.

Multiple integral on a complicated shape

In high-dimensional space, we might want to consider the volume bounded by a function in a complicated shape, say $\mathscr{A}$, that cannot be represented as a union of hyper-rectangles.

Hence, we want to define the volume bounded by a function on a general set $\mathscr{A}$. We can do it by multiplying the value of the function by zero everywhere outside of $\mathscr{A}$ as follows.

Definition

Let $D = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_m, b_m]$ be a m-dimensional hyper-rectangle.
For a general subset $\mathscr{A} \subset D$, we define the multiple integral of f on $\mathscr{A}$ by

$$\int_{\mathscr{A}} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} := \int_D 1_{\mathscr{A}}(\boldsymbol{x}) f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \tag{67}$$

where the indicator function $1_{\mathscr{A}}$ is defined by $1_{\mathscr{A}}(\boldsymbol{x}) := \begin{cases} 1 & \text{if } \boldsymbol{x} \in \mathscr{A}, \\ 0 & \text{if } \boldsymbol{x} \notin \mathscr{A}. \end{cases}$

Probability on a complicated shape

The probability density function can be applied to a complicated shape.

Theorem

Let $X_1, X_2, \ldots, X_m$ be jointly continuous RVs, and let $p_{X_1, X_2, \ldots, X_m}$ be the joint PDF.
For $\mathscr{A} \in \mathbb{R}^m$, assume that it is bounded, i.e., there exists hyper-rectangle $D = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_m, b_m]$ such that $\mathscr{A} \in D$.
Then, we have that

$$\Pr((X_1, X_2, \ldots, X_m) \in \mathscr{A}) = \int_{\mathscr{A}} p_{X_1, X_2, \ldots, X_m}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \tag{68}$$

The assumption about the boundedness of $\mathscr{A}$ will be removed later using improper integrations.

© SUZUKI, Atsushi

We can calculating a multiple integral by the iterated integral

We need to calculate a multiple integral to evaluate the probability of an event related to jointly continuous RVs. How can we do that?

Actually, we can calculate a multiple integral by the ***iterated integral***, according to Fubini-Tonelli Theorem.

Theorem (Fubini-Tonelli Theorem)

Let $f : \mathbb{R}^m \to \mathbb{R}$ be a function and $\mathscr{A} \subset \mathbb{R}^m$ be a subset of $\mathbb{R}^m$ and suppose that there exists a bounded m-dimensional hyper-rectangle $D = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_m, b_m]$.
Then, under a certain loose conditions, we have that

$$\int_{\mathscr{A}} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{a_m}^{b_m} \cdots \left(\int_{a_2}^{b_2} \left(\int_{a_1}^{b_1} 1_{\mathscr{A}}(\boldsymbol{x}) f(\boldsymbol{x})\,\mathrm{d}x_1 \right) \mathrm{d}x_2 \right) \cdots \mathrm{d}x_m \,. \tag{69}$$

Note that the order of the indices is exchangeable.

Special case: calculating a double integral

A bivariable multiple integral is called a ***double integral***. The formula for a double integral is given as follows.

Corollary (Fubini-Tonelli theorem on a double integral)

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a function and $\mathscr{A} \subset \mathbb{R}^2$ be a subset of $\mathbb{R}^2$ and suppose that there exists a bounded 2-dimensional hyper-rectangle $D = [a_1, b_1] \times [a_2, b_2]$. Then, under a certain loose conditions, we have that

$$\begin{aligned} \iint_{\mathscr{A}} f(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 &= \int_{a_2}^{b_2} \left[\int_{a_1}^{b_1} 1_{\mathscr{A}}(x_1, x_2) f(x_1, x_2)\, \mathrm{d}x_1 \right] \mathrm{d}x_2 \\ &= \int_{a_1}^{b_1} \left[\int_{a_2}^{b_2} 1_{\mathscr{A}}(x_1, x_2) f(x_1, x_2)\, \mathrm{d}x_2 \right] \mathrm{d}x_1 . \end{aligned} \tag{70}$$

Strictly speaking, the following condition must be satisfied for the Fubini-Tonelli theorem to hold, i.e., for the iterated integral to give the correct value of the multiple integral.

Condition: The following limit converges (note the absolute value operation).

$$\int_{a_m}^{b_m} \cdots \left(\int_{a_2}^{b_2} \left(\int_{a_1}^{b_1} 1_{\mathscr{A}}(\boldsymbol{x}) |f(\boldsymbol{x})| \, \mathrm{d}x_1 \right) \mathrm{d}x_2 \right) \cdots \mathrm{d}x_m \,. \tag{71}$$

However, the above is rarely an issue in engineering or computer science.

Improper multiple integral

We might want to evaluate the volume bounded by a function's graph in a unbounded domain $\mathscr{A}$. In that case, we define the ***improper multiple integral*** as follows.

Definition (Improper multiple integral)

Let $\mathscr{A} \subset \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}$ be a function defined on $\mathbb{R}^m$. Denote by $\lim_{\substack{a \to -\infty \\ b \to +\infty}}$ the iterated limit operator $\lim_{b \to +\infty} \lim_{a \to -\infty}$. Assume that a certain loose condition is satisfied. Then, we define $\int_{\mathscr{A}} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$ by

$$\int_{\mathscr{A}} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} := \lim_{\substack{a_m \to -\infty \\ b_m \to +\infty}} \cdots \lim_{\substack{a_2 \to -\infty \\ b_2 \to +\infty}} \lim_{\substack{a_1 \to -\infty \\ b_1 \to +\infty}} \int_D 1_{\mathscr{A}}(\boldsymbol{x}) f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \tag{72}$$

where $D = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_m, b_m]$.

Advanced: the loose condition

We assumed some condition in the previous slide. This is because, in fact, the definition in the previous slide is not standard and we usually use another definition for the integral of a function on $\mathbb{R}^m$.

However, if a condition is satisfied, the two definitions are consistent, which is why we imposed the condition. The condition is as follows:

Condition: The following limit converges (note the absolute value operation).

$$\lim_{\substack{a_m \to -\infty \\ b_m \to +\infty}} \int_{a_m}^{b_m} \cdots \left(\lim_{\substack{a_2 \to -\infty \\ b_2 \to +\infty}} \int_{a_2}^{b_2} \left(\lim_{\substack{a_1 \to -\infty \\ b_1 \to +\infty}} \int_{a_1}^{b_1} 1_{\mathscr{A}}(\boldsymbol{x}) |f(\boldsymbol{x})| \,\mathrm{d}x_1 \right) \mathrm{d}x_2 \right) \cdots \mathrm{d}x_m \,. \tag{73}$$

To prove that these two are equivalent, first we define it in the standard way based on the Lebesgue integral, and use the dominant convergence theorem and Fubini-Tonelli's theorem iteratively.

Calculating an improper multiple integral

We can calculate a improper multiple integral by an iterated improper integral.

Theorem (Calculating an improper multiple integral)

Let $\mathscr{A} \subset \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}$ be a function defined on $\mathbb{R}^m$. Assume that the loose condition in the previous slide is satisfied. Then, we have that

$$\int_{\mathscr{A}} f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \int_{-\infty}^{+\infty} \cdots \left(\int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} 1_{\mathscr{A}}(\boldsymbol{x}) f(\boldsymbol{x})\, \mathrm{d}x_1 \right) \mathrm{d}x_2 \right) \cdots \mathrm{d}x_m \,. \tag{74}$$

Special case: an improper multiple integral on the whole space

By substituting $\mathscr{A}$ with $\mathbb{R}^m$, we can define and calculate the improper multiple integral on the whole space $\mathbb{R}^m$. Here, what we need to do is to substitute $1_{\mathbb{R}^m}(\boldsymbol{x}) = 1$ for any $\boldsymbol{x} \in \mathbb{R}^m$.

Corollary (Calculating an improper multiple integral on the whole space)

Let $f : \mathbb{R}^m \to \mathbb{R}$ be a function defined on $\mathbb{R}^m$. Assume that the loose condition is satisfied. Then, we have that

$$\begin{aligned} \int_{\mathbb{R}^m} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &:= \lim_{\substack{a_m \to -\infty \\ b_m \to +\infty}} \cdots \lim_{\substack{a_2 \to -\infty \\ b_2 \to +\infty}} \lim_{\substack{a_1 \to -\infty \\ b_1 \to +\infty}} \int_D 1_{\mathbb{R}^m}(\boldsymbol{x}) f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\ &= \int_{-\infty}^{+\infty} \cdots \left(\int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} f(\boldsymbol{x})\,\mathrm{d}x_1 \right) \mathrm{d}x_2 \right) \cdots \mathrm{d}x_m . \end{aligned} \tag{75}$$

Special case: an improper double integral

Corollary (Calculating an improper double integral)

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a function and $\mathscr{A} \subset \mathbb{R}^2$ be a subset of $\mathbb{R}^2$. Then, under the loose condition, we have that

$$\begin{aligned} \iint_{\mathscr{A}} f(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 &= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} 1_{\mathscr{A}}(x_1, x_2) f(x_1, x_2)\, \mathrm{d}x_1 \right) \mathrm{d}x_2 \\ &= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} 1_{\mathscr{A}}(x_1, x_2) f(x_1, x_2)\, \mathrm{d}x_2 \right) \mathrm{d}x_1 . \end{aligned} \tag{76}$$

© SUZUKI, Atsushi

Steps to calculate a double integral

Recall that we have

$$\begin{aligned}\iint_{\mathscr{A}} f(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 &= \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} 1_{\mathscr{A}}(x_1,x_2) f(x_1,x_2)\,\mathrm{d}x_1\right)\mathrm{d}x_2 \\ &= \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} 1_{\mathscr{A}}(x_1,x_2) f(x_1,x_2)\,\mathrm{d}x_2\right)\mathrm{d}x_1\,.\end{aligned} \tag{77}$$

Hence, we can calculate the double integral $\iint_{\mathscr{A}} f(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2$ as follows.

- **Step 1.**

- **Step 2.**

Steps to calculate a double integral

Recall that we have

$$\iint_{\mathscr{A}} f(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} 1_{\mathscr{A}}(x_1, x_2) f(x_1, x_2)\,\mathrm{d}x_1\right)\mathrm{d}x_2 \\ = \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} 1_{\mathscr{A}}(x_1, x_2) f(x_1, x_2)\,\mathrm{d}x_2\right)\mathrm{d}x_1\,. \tag{77}$$

Hence, we can calculate the double integral $\iint_{\mathscr{A}} f(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2$ as follows.

- **Step 1.** Find the function $g(x_2) := \int_{-\infty}^{+\infty} 1_{\mathscr{A}}(x_1, x_2) f(x_1, x_2)\,\mathrm{d}x_1$ by the improper integral with respect to x_1.
- **Step 2.**

Steps to calculate a double integral

Recall that we have

$$\iint_{\mathscr{A}} f(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} 1_{\mathscr{A}}(x_1,x_2)f(x_1,x_2)\,\mathrm{d}x_1\right)\mathrm{d}x_2$$
$$= \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} 1_{\mathscr{A}}(x_1,x_2)f(x_1,x_2)\,\mathrm{d}x_2\right)\mathrm{d}x_1\,. \tag{77}$$

Hence, we can calculate the double integral $\iint_{\mathscr{A}} f(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2$ as follows.

- **Step 1.** Find the function $g(x_2) := \int_{-\infty}^{+\infty} 1_{\mathscr{A}}(x_1,x_2)f(x_1,x_2)\,\mathrm{d}x_1$ by the improper integral with respect to x_1.
- **Step 2.** Evaluate the improper integral $\int_{-\infty}^{+\infty} g(x_2)\,\mathrm{d}x_2$ with respect to x_2.

Joint PDF example 1: Uniform distribution

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

$$p_{X,Y}(x,y) = \begin{cases} \frac{1}{6} & \text{if } (x,y) \in [0,3] \times [-1,+1], \\ 0 & \text{if } (x,y) \notin [0,3] \times [-1,+1]. \end{cases} \tag{78}$$

For example, the probability $\Pr\left((X,Y) \in [0, \frac{1}{2}] \times [0, \frac{1}{4}]\right)$ is given by

$$\int_0^{\frac{1}{4}} \int_0^{\frac{1}{2}} \frac{1}{6} \, \mathrm{d}x \, \mathrm{d}y = \int_0^{\frac{1}{4}} \left[\frac{1}{6} x \right]_0^{\frac{1}{2}} \mathrm{d}y = \int_0^{\frac{1}{4}} \frac{1}{12} \, \mathrm{d}y = \left[\frac{1}{12} y \right]_0^{\frac{1}{4}} = \frac{1}{48} \tag{79}$$

© SUZUKI, Atsushi

Joint PDF example 2

Example

Let X_1, X_2 be jointly continuous RVs and assume that its joint PDF p_{X_1,X_2} is given by

$$p_{X_1,X_2}(x_1,x_2) = 1_{\mathscr{A}}(x_1,x_2) f(x_1,x_2), \tag{80}$$

where $f(x_1,x_2) = 3 - 3x_1 - \frac{3}{2}x_2$, and $\mathscr{A} = \left\{(x_1,x_2) \in \mathbb{R}^2 \middle| x_1 \geq 0, x_2 \geq 0, \frac{x_1}{1} + \frac{x_2}{2} \leq 1\right\}$.

In the following, we will first confirm that $\iint_{\mathbb{R}^2} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = 1$, then calculate the probability $\Pr((X,Y) \in \mathscr{B})$, where $\mathscr{B} = \left\{(x_1,x_2) \in \mathbb{R}^2 \middle| x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq 1\right\}$.

© SUZUKI, Atsushi

Joint PDF example 2: (i) Confirming the integral on $\mathbb{R}^2$ is 1

Let's calculate $\int_{\mathbb{R}^2} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = 1$ by the iterated integration. We have that $\int_{\mathbb{R}^2} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2\right)\mathrm{d}x_1$. We first evaluate the integral $\int_{-\infty}^{+\infty} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2$.

Joint PDF example 2: (i) Confirming the integral on $\mathbb{R}^2$ is 1

Let's calculate $\int_{\mathbb{R}^2} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = 1$ by the iterated integration. We have that $\int_{\mathbb{R}^2} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2\right)\mathrm{d}x_1$. We first evaluate the integral $\int_{-\infty}^{+\infty} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2$.

Since $p_{X_1,X_2}(x_1,x_2) = 1_{\mathscr{A}}(x_1,x_2) f(x_1,x_2)$, we have that

$$\int_{-\infty}^{+\infty} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2 = \begin{cases} \int_0^{2-2x_1} f(x_1,x_2)\,\mathrm{d}x_2 & \text{if } 0 \le x_1 \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{81}$$

Joint PDF example 2: (i) Confirming the integral on $\mathbb{R}^2$ is 1

Let's calculate $\int_{\mathbb{R}^2} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = 1$ by the iterated integration. We have that $\int_{\mathbb{R}^2} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2\right)\mathrm{d}x_1$. We first evaluate the integral $\int_{-\infty}^{+\infty} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2$.

Since $p_{X_1,X_2}(x_1,x_2) = 1_{\mathscr{A}}(x_1,x_2) f(x_1,x_2)$, we have that

$$\int_{-\infty}^{+\infty} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2 = \begin{cases} \int_0^{2-2x_1} f(x_1,x_2)\,\mathrm{d}x_2 & \text{if } 0 \le x_1 \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{81}$$

Hence, we have that

$$\int_{\mathbb{R}^2} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_0^1 \left(\int_0^{2-2x_1} f(x_1,x_2)\,\mathrm{d}x_2\right)\mathrm{d}x_1 = 1. \tag{82}$$

Joint PDF example 2 (ii) Probability in Region $\mathscr{B}$

Let's calculate $\Pr((x_1, x_2) \in \mathscr{B}) = \int_{\mathbb{R}^2} 1_{\mathscr{B}}(x_1, x_2) p_{X_1, X_2}(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 = 1$ by the iterated integration. We have that $\int_{\mathbb{R}^2} 1_{\mathscr{B}}(x_1, x_2) p_{X_1, X_2}(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 = \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} 1_{\mathscr{B}}(x_1, x_2) p_{X_1, X_2}(x_1, x_2)\, \mathrm{d}x_2 \right) \mathrm{d}x_1$. We first evaluate the integral $\int_{-\infty}^{+\infty} 1_{\mathscr{B}}(x_1, x_2) p_{X_1, X_2}(x_1, x_2)\, \mathrm{d}x_2$.

Joint PDF example 2 (ii) Probability in Region $\mathscr{B}$

Let's calculate $\Pr((x_1,x_2)\in\mathscr{B}) = \int_{\mathbb{R}^2} 1_{\mathscr{B}}(x_1,x_2)p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = 1$ by the iterated integration. We have that $\int_{\mathbb{R}^2} 1_{\mathscr{B}}(x_1,x_2)p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} 1_{\mathscr{B}}(x_1,x_2)p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2\right)\mathrm{d}x_1$. We first evaluate the integral $\int_{-\infty}^{+\infty} 1_{\mathscr{B}}(x_1,x_2)p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2$.

As $\mathscr{B}\subset\mathscr{A}$, we have that $1_{\mathscr{B}}(x_1,x_2)p_{X_1,X_2}(x_1,x_2) = 1_{\mathscr{B}}(x_1,x_2)f(x_1,x_2)$.

Joint PDF example 2 (ii) Probability in Region $\mathscr{B}$

Let's calculate $\Pr((x_1,x_2)\in\mathscr{B}) = \int_{\mathbb{R}^2} 1_{\mathscr{B}}(x_1,x_2) p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = 1$ by the iterated integration. We have that $\int_{\mathbb{R}^2} 1_{\mathscr{B}}(x_1,x_2) p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} 1_{\mathscr{B}}(x_1,x_2) p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2\right)\mathrm{d}x_1$. We first evaluate the integral $\int_{-\infty}^{+\infty} 1_{\mathscr{B}}(x_1,x_2) p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2$.

As $\mathscr{B}\subset\mathscr{A}$, we have that $1_{\mathscr{B}}(x_1,x_2) p_{X_1,X_2}(x_1,x_2) = 1_{\mathscr{B}}(x_1,x_2) f(x_1,x_2)$. Hence

$$\int_{-\infty}^{+\infty} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_2 = \begin{cases} \int_0^{1-x_1} f(x_1,x_2)\,\mathrm{d}x_2 & \text{if } 0 \le x_1 \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{83}$$

Hence, we have that

$$\int_{\mathbb{R}^2} p_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_0^1 \left(\int_0^{1-x_1} f(x_1,x_2)\,\mathrm{d}x_2\right)\mathrm{d}x_1 = \frac{3}{4}. \tag{84}$$

Integration by substitution for a multiple integral

When we want to evaluate a multiple integral of a complicatedly composed function, an integration by substitution might help, as it does for univariate case.

Theorem (Integration by substitution for a multiple integral)

Let $f : \mathbb{R}^m \to \mathbb{R}$ be a m-variable real-valued function and $\boldsymbol{\varphi} : \mathbb{R}^m \to \mathbb{R}^m$ be a bijective differentiable m-variable m-dimensional-vector-valued function. Also, let U be a subset of $\mathbb{R}^m$. Then we have the following.

$$\int_U f(\boldsymbol{\varphi}(\boldsymbol{u})) \left| \det\left(\frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{u}}(\boldsymbol{u}) \right) \right| \mathrm{d}\boldsymbol{u} = \int_{\boldsymbol{\varphi}(U)} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{85}$$

Here, $\det$ indicates the determinant, and $\frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{u}}$ is the Jacobian of $\boldsymbol{\varphi}$.

Difference between a univariable integral and a multiple integral

Strictly, the previous slide's formula is not a strict extension of the univariable case since we have the absolute value operator outside the determinant of the Jacobian. This difference comes because we do not care the direction of the integral as we did in a univariable case.

Specifically, we distinguished $\int_a^b$ and $\int_b^a$ in the univariable case, but we do not care such differences in a multiple integral.

If you want to distinguish them in multiple integral, you can learn a ***differential form*** or ***volume form***.

© SUZUKI, Atsushi

Integration by substitution for a double integral

To see the formula in detail, let us consider the bivariable case.

Corollary (Integration by substitution for a double integral)

$$\begin{aligned}&\int_U f(\varphi_1(u_1,u_2),\varphi_2(u_1,u_2))\left|\det\left(\begin{bmatrix}\frac{\partial\varphi_1}{\partial u_1}(u_1,u_2) & \frac{\partial\varphi_1}{\partial u_2}(u_1,u_2)\\ \frac{\partial\varphi_2}{\partial u_1}(u_1,u_2) & \frac{\partial\varphi_2}{\partial u_2}(u_1,u_2)\end{bmatrix}\right)\right|\mathrm{d}u_1\,\mathrm{d}u_2\\ &=\int_{\boldsymbol{\varphi}(U)} f(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2\end{aligned} \tag{86}$$

Here, recall that

$$\det\left(\begin{bmatrix}a & b\\ c & d\end{bmatrix}\right)=ad-bc. \tag{87}$$

© SUZUKI, Atsushi

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

$$\det\left(\begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix}\right) = \det\left(\begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}\right) = r\cos^2\theta - \left(-r\sin^2\right) = r. \tag{88}$$

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

$$\begin{aligned} \iint_{x^2+y^2\le 1} \left(1 - \sqrt{x^2+y^2}\right) \mathrm{d}x\,\mathrm{d}y &= \int_0^{2\pi}\int_0^1 (1-r)\left|\det\left(\begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix}\right)\right| \mathrm{d}r\,\mathrm{d}\theta \\ &= \int_0^{2\pi}\int_0^1 (1-r)|r|\,\mathrm{d}r\,\mathrm{d}\theta \\ &= \int_0^{2\pi}\left[\int_0^1 \left(r - r^2\right)\mathrm{d}r\right]\mathrm{d}\theta = \int_0^{2\pi}\frac{1}{6}\,\mathrm{d}\theta = \frac{1}{3}\pi. \end{aligned} \tag{89}$$

Outline

3 **Continuous Random Variables**

- Relation among jointly continuous RVs

Note

In the following, we focus on two variable cases to make the discussion easier.
Nonetheless, the same discussion holds for general cases.

© SUZUKI, Atsushi

Marginal PDF (bivariable cases)

We first discuss the PDF of each RV, which helps us see the conditional distribution later, as we did in discrete cases.

Marginal PDF (bivariable cases)

We first discuss the PDF of each RV, which helps us see the conditional distribution later, as we did in discrete cases.

When we have the joint PDF of (X, Y), each of X and Y also has a PDF. To distinguish it from the joint PDF, we call each ***marginal probability density function (marginal PDF)***. We can obtain the explicit form of each by the integral as follows.

Marginal PDF (bivariable cases)

We first discuss the PDF of each RV, which helps us see the conditional distribution later, as we did in discrete cases.

When we have the joint PDF of (X,Y), each of X and Y also has a PDF. To distinguish it from the joint PDF, we call each ***marginal probability density function (marginal PDF)***. We can obtain the explicit form of each by the integral as follows.

Theorem

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

$$
\begin{aligned}
p_X(x) &= \int_{-\infty}^{+\infty} p_{X,Y}(x,y)\mathrm{d}y, \\
p_Y(y) &= \int_{-\infty}^{+\infty} p_{X,Y}(x,y)\mathrm{d}x.
\end{aligned}
\tag{90}
$$

respectively.

© SUZUKI, Atsushi

Conditional PDF (bivariate cases)

Similar to the conditional PMF, we can consider the PDF of a RV updated by knowing the value of the other RV. The updated PDF is called the ***conditional probability distribution function (conditional PDF)***. As in the conditional PMF, the conditional PDF is proportional to the joint PDF. Since the integral of the conditional PDF on the whole real number line must be 1, the conditional PDF is defined as the conditional PDF over the marginal PDF.

Definition

Suppose that (X,Y) is a bivariate continuous RV and its joint PDF is $p_{X,Y}$. Then, for all y such that $p_Y(y) \neq 0$, the ***conditional probability distribution function (conditional PDF)*** $p_{X|Y}$ of X given $Y=y$ is defined by

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}. \tag{91}$$

Note: if $p_Y(y)=0$, the above fraction diverges. However, we do not care it since Y cannot such a value y.

Independence

Similar to discrete RV cases, if the conditional PDF is always the same as the marginal PDF, we say that the two RVs are ***independent***, that is, not related.

Definition (Independence of continuous RVs)

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

When we quantify the relation between RVs, we often calculate the expectation of a function, as we do to evaluate the covariance. We can calculate it using the joint PDF as follows.

Theorem (Expectation of a function of jointly continuous RVs)

Let $(X_1, X_2, \ldots, X_m)$ be a multivariate RV 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

$$\int_{\mathbb{R}^m} f(\boldsymbol{x}) p_{X_1, X_2, \ldots, X_m}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{92}$$

Covariance

For two RVs X and Y, the covariance $\mathrm{Cov}(X,Y)$ is defined by $\mathrm{Cov}(X,Y) := \mathbb{E}\left(X-\mu_X\right)\left(Y-\mu_Y\right)$. We can calculate it using the joint PDF.

Theorem

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

$$\mathrm{Cov}(X,Y) = \iint_{\mathbb{R}^2} \left(x-\mu_X\right)\left(y-\mu_Y\right) p_{X,Y}(x,y) \mathrm{d}x \mathrm{d}y. \tag{93}$$

Example: Multivariate normal distribution

Example (Multivariate normal distribution)

Let $\boldsymbol{\mu}$ be a real m-dimensional vector and $\boldsymbol{\Sigma}$ be a real $m \times m$ positive definite matrix, i.e., a $m \times m$ matrix such that $\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} > 0$ for any non-zero m-dimensional vector $\boldsymbol{x}$. We call the distribution of a m-tuple $(X_1, X_2, \ldots, X_m)$ of RVs a ***multivariate normal distribution*** if it has the following joint PDF $p_{X,Y}$.

$$p_{X_1, X_2, \ldots, X_m}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^m \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right) \tag{94}$$

Example: Multivariate normal distribution

Example (Multivariate normal distribution)

Recall that the joint PDF of a multivariate normal distribution is given as follows.

$$p_{X_1,X_2,\ldots,X_m}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^m \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{94}$$

For a bivariable case $m = 2$, the joint PDF is given by

$$p_{X_1,X_2}(x_1,x_2) = \frac{1}{\sqrt{(2\pi)^2 \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} x_1 & x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix}\right)\begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}^{-1}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)\right), \tag{95}$$

where $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ is a 2-dimensional vector and $\boldsymbol{\Sigma} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}$ is a real 2×2 positive definite matrix.

Example: Multivariate normal distribution

Example (Multivariate normal distribution)

Recall that the joint PDF of a multivariate normal distribution is given as follows.

$$p_{X_1,X_2,\ldots,X_m}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^m \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right) \tag{94}$$

We can see that $p_{X,Y}(x,y)$ takes its maximum if $s = \boldsymbol{\mu}$ since $\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)$ is zero if $s = \boldsymbol{\mu}$ and positive otherwise, according to the positive definite assumption on $\boldsymbol{\Sigma}$.

Example: Multivariate normal distribution

Example (Multivariate normal distribution)

Recall that the joint PDF of a multivariate normal distribution is given as follows.

$$p_{X_1,X_2,\dots,X_m}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^m \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{94}$$

We can see that $p_{X,Y}(x,y)$ takes its maximum if $s = \boldsymbol{\mu}$ since $(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ is zero if $s = \boldsymbol{\mu}$ and positive otherwise, according to the positive definite assumption on $\boldsymbol{\Sigma}$. Unfortunately, we cannot calculate the probability $\Pr((X,Y) \in \mathscr{A})$ analytically for general $\mathscr{A}$.

Example: Multivariate normal distribution

Example (Multivariate normal distribution)

Recall that the joint PDF of a multivariate normal distribution is given as follows.

$$p_{X_1,X_2,\ldots,X_m}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^m \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{94}$$

We can see that $p_{X,Y}(x,y)$ takes its maximum if $s = \boldsymbol{\mu}$ since $(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ is zero if $s = \boldsymbol{\mu}$ and positive otherwise, according to the positive definite assumption on $\boldsymbol{\Sigma}$. Unfortunately, we cannot calculate the probability $\Pr((X,Y) \in \mathscr{A})$ analytically for general $\mathscr{A}$. Nevertheless, we can prove that the mean $\mathbb{E}X_i$ of the ith RV X_i is μ_i, the ith element of the vector $\boldsymbol{\mu}$. Also, the covariance matrix is given by $\boldsymbol{\Sigma}$. In other words, the covariance between $\mathrm{Cov}\left(X_i, X_j\right)$ is given by the entry s_{ij} in the ith row and the jth column of the matrix $\boldsymbol{\Sigma}$. In particular, the variance $\mathbb{V}(X_i) = s_{ii}$.

Example: Multivariate normal distribution

Example (Multivariate normal distribution)

Recall that the joint PDF of a multivariate normal distribution is given as follows.

$$p_{X_1,X_2,\ldots,X_m}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^m \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{94}$$

Suppose that $\boldsymbol{\Sigma}$ is a diagonal matrix, i.e., $s_{ij}=0$ if $i \neq j$. Then, $\det(\boldsymbol{\Sigma}) = \prod_{i=1}^{m} s_i$ and $-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = -\frac{1}{2}\sum_{i=1}^{m}\frac{(x_i-\mu_i)^2}{s_{ii}}$. Therefore, we have the decomposition: $p_{X_1,X_2,\ldots,X_m}(\boldsymbol{x}) = \prod_{i=1}^{m}\frac{1}{\sqrt{2\pi s_{ii}}}\exp\left(-\frac{(x_i-\mu_i)^2}{2s_{ii}}\right)$. Hence, if $\boldsymbol{\Sigma}$ is diagonal, then $X_1,X_2,\ldots,X_m$ are mutually independent.

Outline

3 **Continuous Random Variables**

- Exercises

Exercise (Continuous random variable)

Let X be a random variable, and let F_X be the cumulative distribution function (CDF) of X, given by

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

(1) Evaluate the probability $\Pr(0.25 \leq X \leq 0.75)$.

(2) F_X is differentiable at all but a finite number of points, and its derivative is X's probability density function (p_X), which can be arbitrary at points of non-differentiability. Evaluate $p_X(0.5)$ and $p_X(3)$.

© SUZUKI, Atsushi

Example answer:

(1) From the definition of the cumulative distribution function,
$\Pr(0.25 \leq X \leq 0.75) = F_X(0.75) - F_X(x) \lim_{x \to 0.25}(x)$. Since F_X is a continuous function,
$\lim_{x \to 0.25} F_X(x) = F_X(0.25)$.

Therefore, $\Pr(0.25 \leq X \leq 0.75) = F_X(0.75) - F_X(0.25) = \frac{1}{4}0.75^2 - \frac{1}{4}0.25^2 = \frac{13}{32}$.

Example answer:

(2) Except at the two points $x = 0, 2$, on all of the real line, the derivative of F_X can be simply calculated using the formula for the derivative of a polynomial $\dfrac{d}{dx}x^n = nx^{n-1}$ as follows:

$$\begin{cases} 0 & \text{if } x < 0, \\ \frac{1}{2}x & \text{if } 0 < x < 2, \\ 0 & \text{if } x > 2. \end{cases}$$

Hence, $p_X(0.5) = \frac{1}{2} \cdot (0.5) = 0.25, p_X(3) = 0.$

© SUZUKI, Atsushi

Example answer:

Note: F_X is, in fact, differentiable at $x = 0$. This can be shown since the value of the left derivative $\lim_{h \nearrow 0} \frac{F_X(0+h) - F_X(0)}{h}$ matches the value of the right derivative $\lim_{h \searrow 0} \frac{F_X(0+h) - F_X(0)}{h}$, both being 0, thus the derivative $\dfrac{d}{dx} F_X(0) = 0$.

On the other hand, F_X is not differentiable at $x = 2$. This is because the value of the left derivative $\lim_{h \nearrow 0} \frac{F_X(2+h) - F_X(2)}{h}$ is 1, and the value of the right derivative $\lim_{h \searrow 0} \frac{F_X(2+h) - F_X(2)}{h}$ is 0, and the two do not match.

© SUZUKI, Atsushi

Exercise (Multivariate normal distribution)

(1) Define integral $K(R) = \int_0^R r \exp\left(-\frac{r^2}{2}\right) dr$. Find $K(2)$.

(2) By the change of variables $\begin{cases} x = r\cos\theta, \\ y = r\sin\theta, \end{cases}$ compute the absolute value of the Jacobian determinant $|\det(\frac{\partial(x,y)}{(r,\theta)})| =$ at $r = 0.5, \theta = \pi$.

Exercise (Multivariate normal distribution)

(3) For $R \geq 0$, evaluate the double integral $I(R) = \iint_{x^2+y^2 \leq R^2} \exp\left(-\frac{x^2+y^2}{2}\right) dxdy$ for $R = 2$.

(4) Evaluate the value of the improper double integral $\int_{\mathbb{R}^2} \exp\left(-\frac{x^2+y^2}{2}\right) dxdy$ as R approaches infinity, i.e., $\lim_{R \to +\infty} I(R)$.

(5) The bivariate improper integral discussed in the above (4) can be decomposed into the product of univariate improper integrals as follows:

$$\int_{\mathbb{R}^2} \exp\left(-\frac{x^2+y^2}{2}\right) dxdy = \int_{\mathbb{R}^2} \exp\left(-\frac{x^2}{2}\right) \exp\left(-\frac{y^2}{2}\right) dxdy = \left(\int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{2}\right) dx\right) \left(\int_{-\infty}^{+\infty} \exp\left(-\frac{y^2}{2}\right) dy\right) = \left(\int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{2}\right) dx\right)^2$$

Evaluate the improper integral $\int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{2}\right) dx$.

Exercise (Multivariate normal distribution)

For (6) - (10), let X, Y be random variables with the joint probability density function $p_{X,Y}$ specified by

$p_{X,Y}(x,y) = c \exp\left(-\frac{x^2+y^2}{2}\right)$,

where c is a constant.

(6) Given that $p_{X,Y}$ is a joint probability density function, determine the constant c.

(7) Calculate the probability $\Pr(X^2 + Y^2 \leq 2^2)$.

(8) The marginal probability density function for X, $p_X(x)$, is found by $p_X(x) = \int_{-\infty}^{+\infty} p_{X,Y}(x,y)dy$. Evaluate $p_X(-2)$.

(9) The conditional probability density function for Y given X, $p_{Y|X}(y|x)$, is calculated by $p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$. Evaluate $p_{Y|X}(0|-2)$.

(10) Evaluate the expected value of $(X^2+Y^2)^2$, $\mathbb{E}[(X^2+Y^2)^2]$.

© SUZUKI, Atsushi

Exercise (Multivariate normal distribution)

For (6) - (10), let X, Y be random variables with the joint probability density function $p_{X,Y}$ specified by

$$p_{X,Y}(x,y) = c \exp\left(-\frac{x^2+y^2}{2}\right),$$

where c is a constant.

(11) Select the **ONE correct statement** from the above:

- X and Y are independent, and the covariance of X and Y is 0. (correct)
- X and Y are independent, and the covariance of X and Y is non-zero.
- X and Y are not independent, and the covariance of X and Y is 0.
- X and Y are not independent, and the covariance of X and Y is non-zero.

© SUZUKI, Atsushi

Example answer:

(1) $K(R) = \int_0^R r \exp\left(-\frac{r^2}{2}\right) dr$ can be calculated using a substitution of variables with $s = r^2$,
leading to $\int_0^R r \exp\left(-\frac{r^2}{2}\right) dr = \int_0^{R^2} \exp(-s) \frac{ds}{dr} dr = \int_0^{R^2} \exp(-s) ds$.
Since $\int_0^{R^2} \exp(-s) ds = [-\exp(-s)]_0^{R^2} = 1 - \exp\left(-R^2\right)$, for $R = 2$, we have $K(2) = 1 - \exp\left(-2^2\right)$.

Example answer:

(2) By definition, the Jacobian is given by $\frac{\partial(x,y)}{\partial(r,\theta)} = \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix}$. The first column vector $\begin{bmatrix} \frac{\partial x}{\partial r} \\ \frac{\partial y}{\partial r} \end{bmatrix}$ represents the velocity vector of the (x,y) coordinates moving at unit speed in the positive direction of r with θ held fixed, and the second column vector $\begin{bmatrix} \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial \theta} \end{bmatrix}$ represents the velocity vector of the (x,y) coordinates when θ is moved at unit speed with r held fixed. Calculating the Jacobian, we find $\begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}$, and thus its determinant is $r(\cos^2\theta + \sin^2\theta) = r$, which is always positive, meaning the absolute value of the determinant of the Jacobian is r (independent of the value of θ). Therefore, for $r = 0.5$, $\theta = \pi$, we have $|\det(\frac{\partial(x,y)}{\partial(r,\theta)})| = 0.5$. This coordinate transformation is known as polar coordinate transformation, where r represents the distance from the point (x,y) to the origin, and θ represents the angle formed by the line segment (0, 0) - (x, y) with the positive direction of the x-axis.

Example answer:

(3) The formula for variable substitution (substitution integration) in double integrals is given by $\int_{A'} f(x,y)dxdy = \int_A f(x(r,\theta),y(r,\theta))|\frac{\partial(x,y)}{\partial(r,\theta)}|drd\theta$, where the right side uses notation loosely, with r and θ representing the functions for x and y values respectively, and A' and A are the regions corresponding through the variable transformation. In this problem, the region A corresponding to $x^2+y^2 \le R^2$ translates in the (r,θ) coordinate system to a region A' satisfying $0 < r \le R$ and $0 \le \theta < 2\pi$, the original domain of θ. Paying attention to the calculated $|\frac{\partial(x,y)}{\partial(r,\theta)}| = r$, we can compute $\int_{x^2+y^2 \le R^2} \exp\left(-\frac{x^2+y^2}{2}\right)dxdy = \int_0^{2\pi} \int_0^R \exp\left(-\frac{r^2}{2}\right)rdrd\theta$. Evaluating the double integral by computing the integral over r first, as done in (1) where $K(R) = 1-\exp\left(-R^2\right)$, we find $I(R) = \int_0^{2\pi}(1-\exp\left(-R^2\right))d\theta = 2\pi(1-\exp\left(-R^2\right))$. Therefore, $I(2) = 2\pi(1-\exp\left(-2^2\right))$.

(4) $\int_{\mathbb{R}^2} \exp\left(-\frac{x^2+y^2}{2}\right)dxdy = \lim_{R\to+\infty} I(R) = \lim_{R\to+\infty} 2\pi(1-\exp\left(-R^2\right)) = 2\pi$.

Example answer:

(5) $2\pi = \int_{\mathbb{R}^2} \exp\left(-\frac{x^2+y^2}{2}\right) dxdy = (\int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{\sqrt{2}}\right) dx)^2$. Since $\exp\left(-\frac{x^2}{\sqrt{2}}\right)$ is always positive, $\int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{\sqrt{2}}\right) dx$ is non-negative. Hence, $\int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{\sqrt{2}}\right) dx = \sqrt{2\pi}$. It is known that the antiderivative of $\exp\left(-\frac{x^2}{\sqrt{2}}\right)$ is not an elementary function, making it difficult to directly compute this improper integral as the limit of a definite integral. This problem approached the double integral and polar coordinate transformation, an idea dating back to Poisson in the 19th century, and this broad integral is known as the Gaussian integral or Euler-Poisson integral.

Example answer:

(6) Since $p_{X,Y}$ is the joint probability density function, $\int_{\mathbb{R}^2} p_{X,Y}(x,y)dxdy = 1$. Using the result from (4), we can compute $\int_{\mathbb{R}^2} p_{X,Y}(x,y)dxdy = \int_{\mathbb{R}^2} c\exp\left(-\frac{x^2+y^2}{2}\right)dxdy = c2\pi$. Therefore, $c = \frac{1}{2\pi}$.

(7) From the definition of the joint probability density function, $\Pr(X^2+Y^2 \leq R^2) = \int_{x^2+y^2\leq R^2} p_{X,Y}(x,y)dxdy$. Using the definition of $p_{X,Y}$ and the value of c found in (6), we have $\int_{x^2+y^2\leq R^2} p_{X,Y}(x,y)dxdy = \int_{x^2+y^2\leq R^2} \frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right)dxdy = \frac{1}{2\pi}I(R)$. Thus, $\Pr(X^2+Y^2 \leq 2^2) = \frac{1}{2\pi}I(2) = \frac{1}{2\pi}2\pi(1-\exp(-2^2))$.

© SUZUKI, Atsushi

Example answer:

(8) The integrand $p_{X,Y}(x,y) = \frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right) = \frac{1}{2\pi}\exp\left(-\frac{x^2}{2}\right)\exp\left(-\frac{y^2}{2}\right)$ can be decomposed into functions of x and y. Utilizing this,

$p_X(x) = \int_{-\infty}^{+\infty} p_{X,Y}(x,y)dy = \int_{-\infty}^{+\infty} \frac{1}{2\pi}\exp\left(-\frac{x^2}{2}\right)\exp\left(-\frac{y^2}{2}\right)dy = \frac{1}{2\pi}\exp\left(-\frac{x^2}{2}\right)\int_{-\infty}^{+\infty}\exp\left(-\frac{y^2}{2}\right)dy.$

From (5), $\int_{-\infty}^{+\infty}\exp\left(-\frac{y^2}{2}\right)dy = \sqrt{2\pi}$ hence, $p_X(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$. Therefore, $p_X(-2) = \frac{e^2}{\sqrt{2\pi}}$.

(9) From $p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$, we have $p_{Y|X}(0|-2) = \frac{\frac{e^2}{2\pi}}{\frac{e^2}{\sqrt{2\pi}}} = \frac{1}{\sqrt{2\pi}}$.

Example answer:

(10) $\mathbb{E}[(X^2+Y^2)^2] = \int_{\mathbb{R}^2}(x^2+y^2)^2 p_{X,Y}(x,y)dxdy$ can be calculated using the joint probability density function. Utilizing polar coordinate transformation similarly to (3) and (4), where $(x^2+y^2)^2 = r^4$, we compute $\int_{\mathbb{R}^2}(x^2+y^2)^2 p_{X,Y}(x,y)dxdy = \lim_{R\to+\infty}\int_0^{2\pi} L(R)d\theta$, where $L(R) = \int_0^R r^4 \frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right)\cdot rdr$. Using the substitution $s = \frac{r^2}{2}$, we find it equals $\int_0^{R^2} 4s^2 \frac{1}{2\pi}\exp(-s)ds$. Evaluating the antiderivative of $s^2\exp(-s)$ through integration by parts twice, we find $\int s^2\exp(-s) = -s^2\exp(-s) - 2s\exp(-s) - 2exp(-s) + C$, where C is the integration constant. Thus, $L(R) = \frac{2}{\pi}(2-2(R^4+2R^2+2)\exp\left(-R^2\right))$. Therefore, $\int_{\mathbb{R}^2}(x^2+y^2)^2 p_{X,Y}(x,y)dxdy = \lim_{R\to+\infty}\int_0^{2\pi} L(R)d\theta = 8$.

(11) From the result of (8), $p_X(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$, and similarly, $p_Y(y) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)$. Thus, $p_{X,Y}(x,y) = p_X(x)p_Y(y)$, and the random variables X and Y are independent. Therefore, the covariance and correlation coefficient between X and Y are 0.

Outline

4 Sample Statistics

- Introduction: why do we learn sample statistics?
- Terminology
- Sample mean, law of large numbers, and central limit theorem
- Estimation of distribution and parametric model
- Likelihood
- Maximum likelihood estimator
- Exercises

© SUZUKI, Atsushi

Outline

4 **Sample Statistics**

- Introduction: why do we learn sample statistics?

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**.

© SUZUKI, Atsushi

Learning outcomes

By the end of this section, 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

4 **Sample Statistics**

- Terminology

Population and sample

In the context of statistics,

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

Summary statistics and sample statistics

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

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

© SUZUKI, Atsushi

Outline

4 **Sample Statistics**

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

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***

$$\overline{X}_m = \frac{1}{m}(X_1 + X_2 + \cdots + X_m), \tag{95}$$

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 $\overline{X}_m$ to approximate the true mean μ, which is unknown.

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

© SUZUKI, Atsushi

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*

$$\overline{X}_m := \frac{1}{m}(X_1 + X_2 + \cdots + X_m). \tag{96}$$

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***.

© SUZUKI, Atsushi

What is the standard normal distribution?

The ***standard normal distribution***, also known as the ***standard Gaussian distribution*** is the distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{97}$$

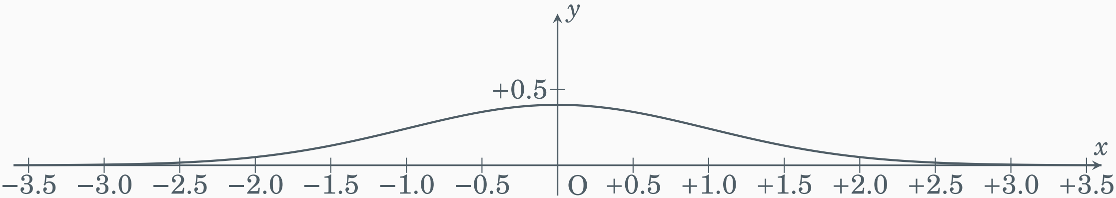


Figure: The standard normal distribution's PDF.

Mean: 0, Variance: 1. The PDF is symmetric about $x = 0$ and it is dense around $x = 0$.

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*

$$\overline{X}_m := \frac{1}{m}(X_1 + X_2 + \cdots + X_m). \tag{98}$$

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

© SUZUKI, Atsushi

The standard normal distribution's CDF

By definition, the CDF $F : \mathbb{R} \to [0, 1]$ of the standard normal distribution is given by

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x'^2}{2}\right) \mathrm{d}x' . \tag{99}$$

It is known that this function is not elementary.

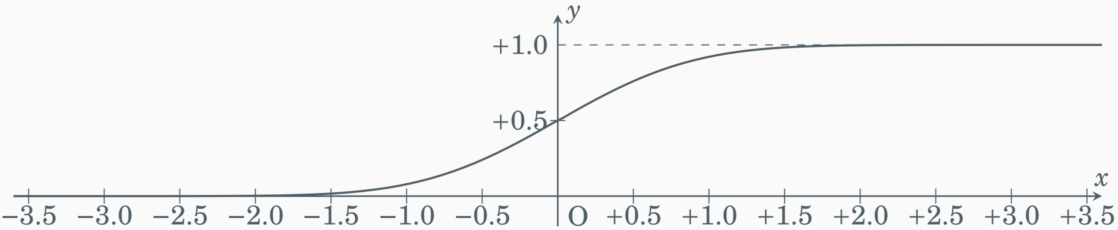


Figure: The standard normal distribution's CDF.

© SUZUKI, Atsushi

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 0 or $+1$ with probability $\frac{1}{2}$ for each.
Then the mean and the variance of X_i are $\frac{1}{2}$ and $\frac{1}{4}$, respectively.
According to the CLT, the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ converges to that of the standard normal distribution $\mathcal{N}(0,1)$.

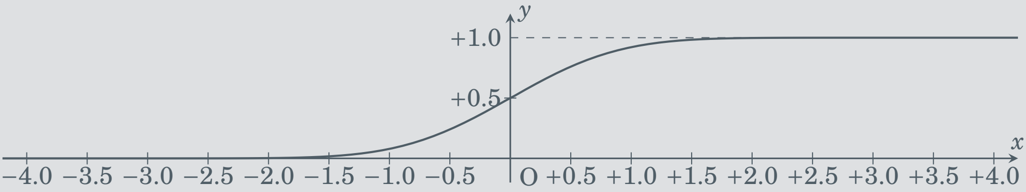


Figure: Dashed: the standard normal distribution's CDF.

Example of the convergence by the CLT

Example

Let $X_1, X_2, \dots$ be an infinite sequence of independently identically distributed RVs, where X_i takes 0 or $+1$ with probability $\frac{1}{2}$ for each.

Then the mean and the variance of X_i are $\frac{1}{2}$ and $\frac{1}{4}$, respectively.

According to the CLT, the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ converges to that of the standard normal distribution $\mathcal{N}(0,1)$.

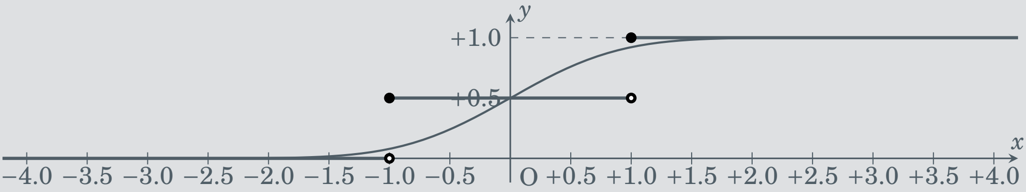


Figure: Dashed: the standard normal distribution's CDF. Solid: the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$, where $m = 1$.

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 0 or $+1$ with probability $\frac{1}{2}$ for each.
Then the mean and the variance of X_i are $\frac{1}{2}$ and $\frac{1}{4}$, respectively.
According to the CLT, the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ converges to that of the standard normal distribution $\mathcal{N}(0, 1)$.

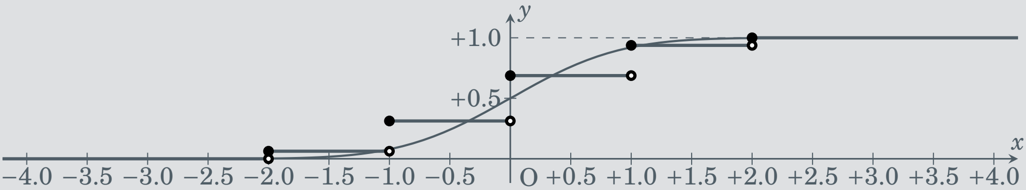


Figure: Dashed: the standard normal distribution's CDF. Solid: the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$, where $m = 4$.

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 0 or $+1$ with probability $\frac{1}{2}$ for each.
Then the mean and the variance of X_i are $\frac{1}{2}$ and $\frac{1}{4}$, respectively.
According to the CLT, the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ converges to that of the standard normal distribution $\mathcal{N}(0,1)$.

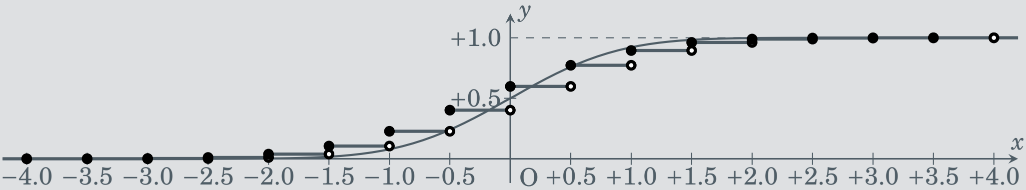


Figure: Dashed: the standard normal distribution's CDF. Solid: the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$, where $m = 16$.

Example of the convergence by the CLT

Example

Let $X_1, X_2, \dots$ be an infinite sequence of independently identically distributed RVs, where X_i takes 0 or $+1$ with probability $\frac{1}{2}$ for each.
Then the mean and the variance of X_i are $\frac{1}{2}$ and $\frac{1}{4}$, respectively.
According to the CLT, the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ converges to that of the standard normal distribution $\mathcal{N}(0,1)$.
Note that the CLT is about the CDF, but **NOT about the PDF**. The convergence of the PDF does not always hold. Specifically, in the above case, $\overline{X_m}$ is a discrete random variable since each X_i is. Hence, the random variable $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ does not have a PDF. Therefore, we **CANNOT** say that the PDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ converges to that of the standard normal distribution.

© SUZUKI, Atsushi

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**.

Outline

4 **Sample Statistics**

- Estimation of distribution and parametric model

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[10], 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.

[10] 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 example 1: Bernoulli distribution

Example (Bernoulli distribution)

The Bernoulli distribution[11] is a discrete parametric model with a sole parameter, which 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 $P(0; \theta) = 1 - \theta$.

Theorem

The mean and the variance of a RV following the Bernoulli distribution with the parameter θ are θ and $\theta(1 - \theta)$, respectively.

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

© SUZUKI, Atsushi

Parametric model example 2: normal distribution

Example (Normal distribution)

The normal distribution, also known as the ***Gaussian distribution***, is a continuous parametric model, which has mean parameter $\mu \in \mathbb{R}$ and variance parameter $\sigma^2 \in \mathbb{R}_{>0}$. 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 $p\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

Theorem

The mean and the variance of a RV following the normal distribution with mean parameter μ and variance parameter σ^2 are μ and σ^2, respectively.

PDF of 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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{100}$$

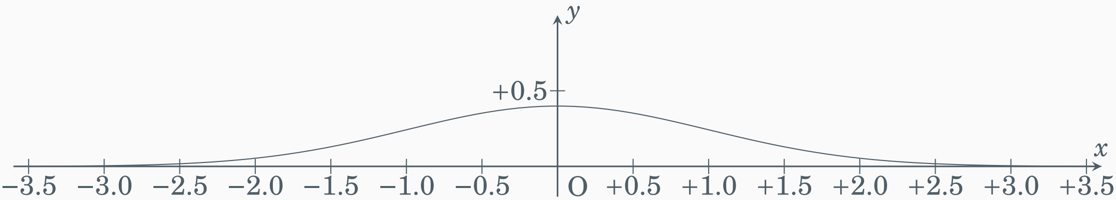


Figure: Normal distributions' PDF ($\mu = 0, \sigma = 1$).

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

PDF of 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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{100}$$

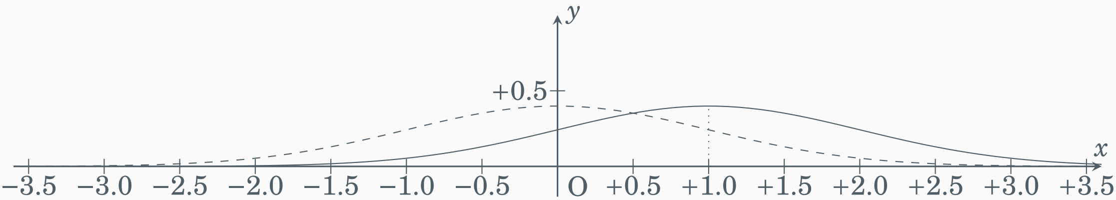


Figure: Normal distributions' PDF (Solid: $\mu = 1, \sigma = 1$, Dashed: $\mu = 0, \sigma = 1$).

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

© SUZUKI, Atsushi

PDF of 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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{100}$$

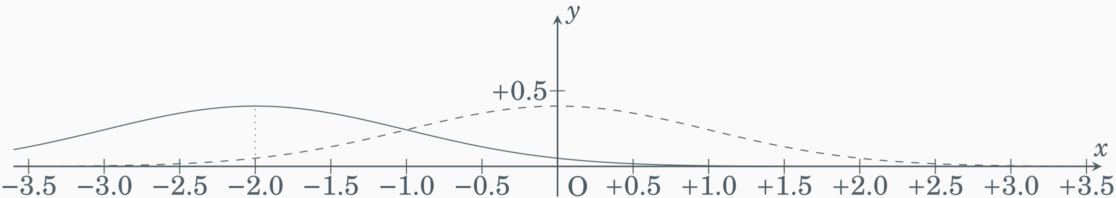


Figure: Normal distributions' PDF (Solid: $\mu = -2, \sigma = 1$, Dashed: $\mu = 0, \sigma = 1$).

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

© SUZUKI, Atsushi

PDF of 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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{100}$$

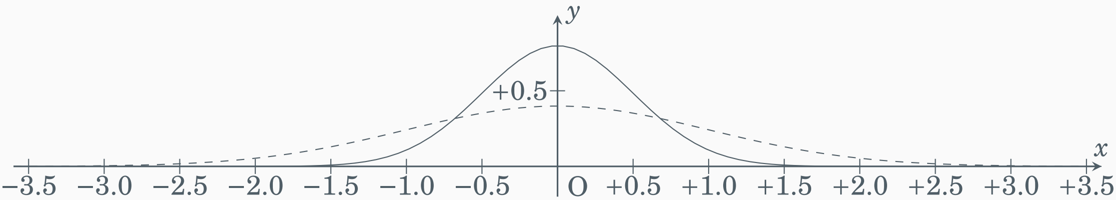


Figure: Normal distributions' PDF (Solid: $\mu = 0, \sigma = 0.5$, Dashed: $\mu = 0, \sigma = 1$).

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

PDF of 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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{100}$$

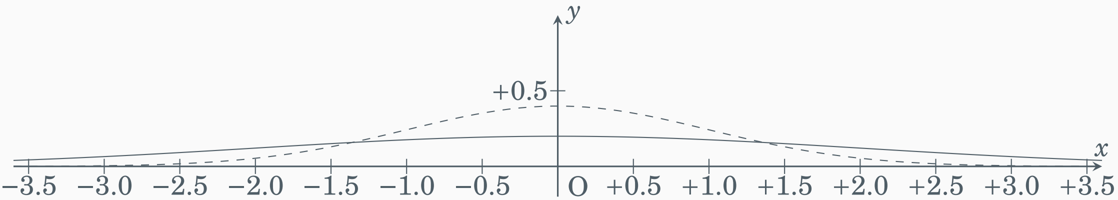


Figure: Normal distributions' PDF (Solid: $\mu = 0, \sigma = 2.0$, Dashed: $\mu = 0, \sigma = 1$).

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

PDF of 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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{100}$$

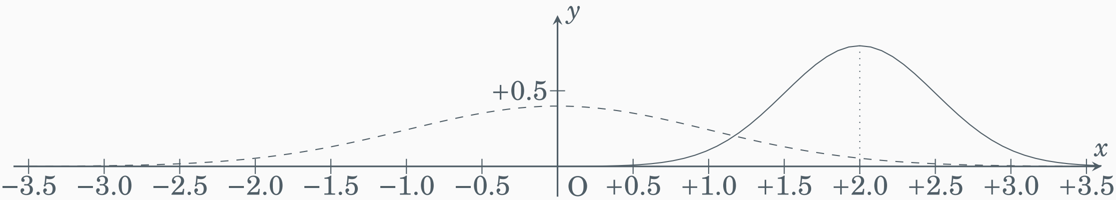


Figure: Normal distributions' PDF (Solid: $\mu = 2, \sigma = 0.5$, Dashed: $\mu = 0, \sigma = 1$).

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

PDF of 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:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{100}$$

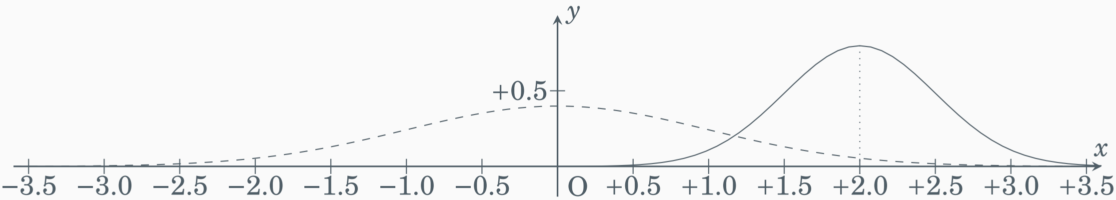


Figure: Normal distributions' PDF (Solid: $\mu = 2, \sigma = 0.5$, Dashed: $\mu = 0, \sigma = 1$).

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

© SUZUKI, Atsushi

Outline

4 Sample Statistics

- Likelihood

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.

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{101}$$

Likelihood

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

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

Definition (Likelihood of a continuous parametric model)

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

$$p(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot p(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot p(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{101}$$

© SUZUKI, Atsushi

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

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot \theta. \tag{102}$$

- 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

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta \cdot \theta \cdot (1 - \theta) \cdot \theta. \qquad (102)$$

- 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

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta\cdot\theta\cdot(1-\theta)\cdot\theta. \tag{102}$$

- 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 $\frac{1}{2}\cdot\frac{1}{2}\cdot\left(1-\frac{1}{2}\right)\cdot\frac{1}{2} = \frac{1}{16} = \frac{16}{256}$.

© SUZUKI, Atsushi

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

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot \theta. \tag{102}$$

- 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

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot \theta. \tag{102}$$

- 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

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta \cdot \theta \cdot (1 - \theta) \cdot \theta. \tag{102}$$

- 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 $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$.

© SUZUKI, Atsushi

Probability and likelihood

The value of the product

$$P(\boldsymbol{x}_1;\boldsymbol{\theta})\cdot P(\boldsymbol{x}_2;\boldsymbol{\theta})\cdot\cdots\cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}) \tag{103}$$

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

4 **Sample Statistics**

- Maximum likelihood estimator

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

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{104}$$

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

© SUZUKI, Atsushi

MLE maximizes the score and minimizes the negative log likelihood

For a parameter vector $\boldsymbol{\theta}$, the following is equivalent[12].

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

$$P(\boldsymbol{x}_1;\boldsymbol{\theta})\cdot P(\boldsymbol{x}_2;\boldsymbol{\theta})\cdot\cdots\cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{105}$$

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

$$\log P(\boldsymbol{x}_1;\boldsymbol{\theta})+\log P(\boldsymbol{x}_2;\boldsymbol{\theta})+\cdots+\log P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{106}$$

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

$$-\log P(\boldsymbol{x}_1;\boldsymbol{\theta})-\log P(\boldsymbol{x}_2;\boldsymbol{\theta})-\cdots-\log P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{107}$$

---

[12] 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.

© SUZUKI, Atsushi

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

$$\begin{aligned}
&\log\left(2\pi\sigma^2\right) + \frac{\left(x_1-\mu\right)^2}{2\sigma^2} + \log\left(2\pi\sigma^2\right) + \frac{\left(x_2-\mu\right)^2}{2\sigma^2} + \cdots + \log\left(2\pi\sigma^2\right) + \frac{\left(x_m-\mu\right)^2}{2\sigma^2} \\
&= m\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left[\left(x_1-\mu\right)^2 + \left(x_2-\mu\right)^2 + \cdots + \left(x_m-\mu\right)^2\right].
\end{aligned} \tag{108}$$

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

$$
\begin{aligned}
&\log\left(2\pi\sigma^2\right) + \frac{\left(x_1-\mu\right)^2}{2\sigma^2} + \log\left(2\pi\sigma^2\right) + \frac{\left(x_2-\mu\right)^2}{2\sigma^2} + \cdots + \log\left(2\pi\sigma^2\right) + \frac{\left(x_m-\mu\right)^2}{2\sigma^2} \\
&= m\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left[\left(x_1-\mu\right)^2 + \left(x_2-\mu\right)^2 + \cdots + \left(x_m-\mu\right)^2\right].
\end{aligned}
\tag{108}
$$

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

$$-\log P(x_1;\theta)P(x_2;\theta)\ldots P(x_m;\theta) = m_0 \log(1-\theta) + m_1 \log\theta, \tag{109}$$

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 $\bar{x} = \frac{m_1}{m}$. Let l denote the above negative log-likelihood. Suppose that $m_0 \neq 0$ and $m_1 \neq 0$, then l takes the minimum[13] if and only if $\theta = \frac{m_1}{m} = \bar{x}$. Hence, the MLE $\hat{\theta} = \frac{m_1}{m} = \bar{x}$.
For example, if $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, then $\hat{\theta} = \frac{m_1}{m} = \bar{x} = \frac{3}{4}$.

[13] 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 as $m \to \infty$.
- ***Asymptotic normality***: An appropriately scaled MLE's distribution converges to a normal distribution, and its error is proportional to $\frac{1}{\sqrt{m}}$ for sufficiently large m.

© SUZUKI, Atsushi

Outline

4 **Sample Statistics**

- Exercises

Exercise (Standard normal distribution)

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

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

Outline

5 **Statistical Test**

- Introduction: why do we learn statistical tests?

Statistical tests support our judgements

In real applications (e.g., physical, engineering, medical, etc.), we need to judge from data whether a phenomenon happens or not.

Specifically, for some summary statistics or parameter θ and a set $\mathscr{H}_1$, we often want to judge from data points whether $\theta \in \mathscr{H}_1$ or not.

Statistical tests support our judgements

In real applications (e.g., physical, engineering, medical, etc.), we need to judge from data whether a phenomenon happens or not.

Specifically, for some summary statistics or parameter θ and a set $\mathscr{H}_1$, we often want to judge from data points whether $\theta \in \mathscr{H}_1$ or not.

For example, if we investigate the purity of a factory's chemical product, we might want to know whether the true expectation μ of the purity is the same as the purity μ_0 of the natural material or not.

In this case, $\mathscr{H}_1 = [0,1] \setminus \{\mu_0\}$, and we want to discuss whether $\theta \in \mathscr{H}_1$ or not.

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

© SUZUKI, Atsushi

Learning outcomes

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

- Explain the logic of statistical tests
- Explain the definitions of p-value, significance level, type-I error, and type-II error.
- Make a judgment from data using statistical tests

Outline

5 Statistical Test

- The logic of statistical tests

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.

A naïve idea is to evaluate the “probability” of $\theta \in \mathscr{H}_1$ when the data points’ values are $x_1, x_2, \ldots, x_m$.

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.

A naïve idea is to evaluate the “probability” of $\theta \in \mathscr{H}_1$ when the data points’ values are $x_1, x_2, \ldots, x_m$.

However, in (frequentism) statistics, we cannot discuss the probability of a parameter θ being in a set since a parameter θ is not a random variable, while it regards data points $x_1, x_2, \ldots, x_m$ as values of random variables.

© SUZUKI, Atsushi

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.

A naïve idea is to evaluate the “probability” of $\theta \in \mathscr{H}_1$ when the data points’ values are $x_1, x_2, \ldots, x_m$.

However, in (frequentism) statistics, we cannot discuss the probability of a parameter θ being in a set since a parameter θ is not a random variable, while it regards data points $x_1, x_2, \ldots, x_m$ as values of random variables.

In contrast, we can discuss the other direction, that is, given a parameter θ, we can discuss the probability of the random variables taking the given values $x_1, x_2, \ldots, x_m$.

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

© SUZUKI, Atsushi

The fundamental logic of statistical test

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 $x_1, x_2, \ldots, x_m$."

Hence, discussing the event $\theta \notin \mathscr{H}_1$ is essential.

The fundamental logic of statistical test

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 $x_1, x_2, \ldots, x_m$."

Hence, discussing the event $\theta \notin \mathscr{H}_1$ is essential.

Let $\mathscr{H}$ be the set of all the possible values that θ can take and define $\mathscr{H}_0 := \mathscr{H} \setminus \mathscr{H}_1$.

The event $\theta \notin \mathscr{H}_1$, which we focus on, is equivalent to $\theta \in \mathscr{H}_0$.

Hence, $\mathscr{H}_0$ plays an essential role in statistical tests. $\mathscr{H}_0$ is called the ***null hypothesis*** and $\mathscr{H}_1$ is called the ***alternative hypothesis***.

In statistical tests, a ***hypothesis*** is a set of values that the variable θ, which we are interested in, may take.

Test statistics

Our starting point is to assume $\theta \notin \mathscr{H}_1$, or equivalently, $\theta \in \mathscr{H}_0$. Our objective is that the data points $x_1, x_2, \ldots, x_m$ “contradict in a probability theory sense” the assumption.

Test statistics

Our starting point is to assume $\theta \notin \mathscr{H}_1$, or equivalently, $\theta \in \mathscr{H}_0$. Our objective is that the data points $x_1, x_2, \ldots, x_m$ “contradict in a probability theory sense” the assumption.

To judge whether a “contradiction” happens, we evaluate a summary statistic of the empirical distribution. Such a summary statistic is called a ***test statistic***. The test statistic is a RV since it is a function of the data points, which are the values of RVs. Hence, the distribution of a test statistic is determined if we fix a distribution of the data points.

© SUZUKI, Atsushi

Test statistics

Our starting point is to assume $\theta \notin \mathscr{H}_1$, or equivalently, $\theta \in \mathscr{H}_0$. Our objective is that the data points $x_1, x_2, \ldots, x_m$ “contradict in a probability theory sense” the assumption.

To judge whether a “contradiction” happens, we evaluate a summary statistic of the empirical distribution. Such a summary statistic is called a ***test statistic***. The test statistic is a RV since it is a function of the data points, which are the values of RVs. Hence, the distribution of a test statistic is determined if we fix a distribution of the data points.

For a distribution corresponding to $\mathscr{H}_0$, if the value of the test statistic is unlikely taken on the distribution (i.e. if a “probabilistic contradiction” happens), then we can conclude that the data points are not generated by the distribution. That is, we can conclude $\theta \notin \mathscr{H}_0$, i.e., $\theta \in \mathscr{H}_1$. This is the basic idea of the statistical test.

© SUZUKI, Atsushi

Terminology: rejecting and accepting a hypothesis

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

© SUZUKI, Atsushi

Outline

5 **Statistical Test**

- Example test statistics

Example: are our products better?

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

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

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

Table: 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?

© SUZUKI, Atsushi

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, owing to its stochastic behavior. This is our concern.

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

t-test about the true expectation

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

We want to see whether or not the true mean equals 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***.

$$t := \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{m}}}, \tag{110}$$

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

$$\overline{X} := \frac{1}{m}\sum_{i=1}^{m} X_i, \quad s := \sqrt{\frac{1}{m}\left(X_i - \overline{X}\right)^2}. \tag{111}$$

© SUZUKI, Atsushi

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.

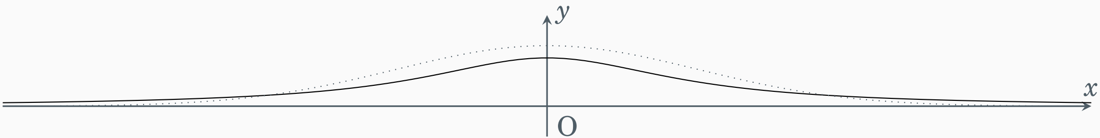


Figure: Black solid curve: the PDF of the t-distribution with 1 degree of freedom.
Black dotted curve: the PDF of the standard normal distribution.

The t-distribution's PDF is symmetric and similar to the standard normal distribution's PDF but has a larger probability of taking extremely large or small values.

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.

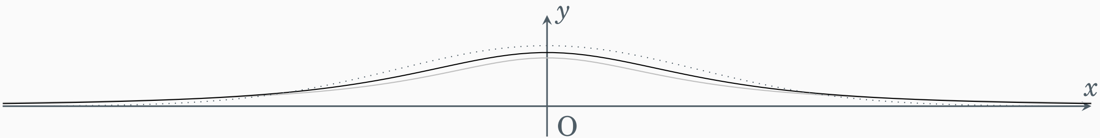


Figure: Black solid curve: the PDF of the t-distribution with 2 degree of freedom.
Black dotted curve: the PDF of the standard normal distribution.

The t-distribution's PDF is symmetric and similar to the standard normal distribution's PDF but has a larger probability of taking extremely large or small values.

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

© SUZUKI, Atsushi

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.

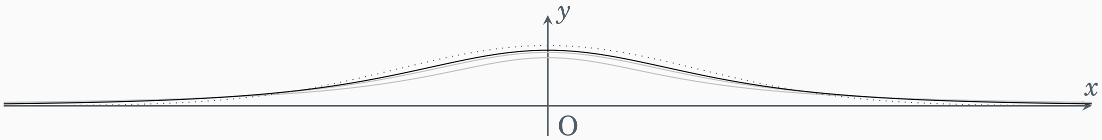


Figure: Black solid curve: the PDF of the t-distribution with 3 degree of freedom.
Black dotted curve: the PDF of the standard normal distribution.

The t-distribution's PDF is symmetric and similar to the standard normal distribution's PDF but has a larger probability of taking extremely large or small values.

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

$$p_{m-1}(x) = \frac{\Gamma\left(\frac{m}{2}\right)}{\sqrt{(m-1)\pi}\,\Gamma\left(\frac{m-1}{2}\right)}\left(1+\frac{x^2}{m-1}\right)^{-\frac{m}{2}} \tag{112}$$

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

t-test is not limited to the one about the true expectation.

We have focused on a statistical test about the true expectation.

In general, a statistic is called a t statistic if it follows the t distribution. Also, a statistical test using a t statistic is called a t test. Hence, if you find a t test in another context, it might not be about the true expectation. It is always essential to confirm what the null hypothesis is and what the alternative hypothesis is in the context you are interested in.

© SUZUKI, Atsushi

Outline

5 **Statistical Test**

- p-value

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

[14] 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.

© SUZUKI, Atsushi

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

$$p = \Pr(|t| > |t_0|). \tag{113}$$

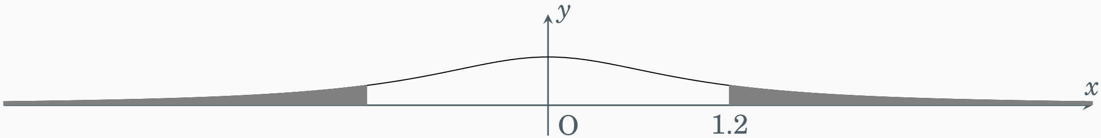


Figure: The p-value (the gray area) when t takes $t_0 = 1.2$.

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

$$p = \Pr(|t| > |t_0|). \tag{113}$$

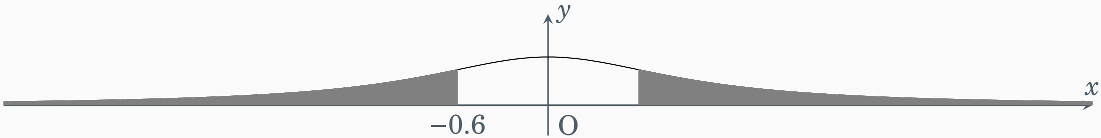


Figure: The p-value (the gray area) when t takes $t_0 = -0.6$.

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

$$p = \Pr(|t| > |t_0|). \tag{113}$$

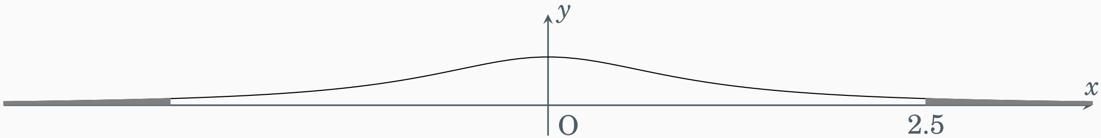


Figure: The p-value (the gray area) when t takes $t_0 = 2.5$.

© SUZUKI, Atsushi

Significance level

We reject a hypothesis consisting of a single distribution if the p-value of the distribution on the data points is small[15].

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 $\mathscr{H}_1$.

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

© SUZUKI, Atsushi

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) and determine which statistic to use.
- **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) and determine which statistic to use. For example, if we are interested in the true expectation, the null hypothesis is $\mu = \mu_0$, where μ is the unknown true expectation and μ_0 is a value, which we decide. The alternative hypothesis is $\mu \neq \mu_0$. We can use t statistic.

- **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) and determine which statistic to use. For example, if we are interested in the true expectation, the null hypothesis is $\mu = \mu_0$, where μ is the unknown true expectation and μ_0 is a value, which we decide. The alternative hypothesis is $\mu \neq \mu_0$. We can use t statistic.
- **Step 2**: Calculate the statistic.
- **Step 3**:
- **Step 4**:

© SUZUKI, Atsushi

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) and determine which statistic to use. For example, if we are interested in the true expectation, the null hypothesis is $\mu = \mu_0$, where μ is the unknown true expectation and μ_0 is a value, which we decide. The alternative hypothesis is $\mu \neq \mu_0$. We can use t statistic.
- **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) and determine which statistic to use. For example, if we are interested in the true expectation, the null hypothesis is $\mu = \mu_0$, where μ is the unknown true expectation and μ_0 is a value, which we decide. The alternative hypothesis is $\mu \neq \mu_0$. We can use t statistic.
- **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**:

© SUZUKI, Atsushi

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) and determine which statistic to use. For example, if we are interested in the true expectation, the null hypothesis is $\mu = \mu_0$, where μ is the unknown true expectation and μ_0 is a value, which we decide. The alternative hypothesis is $\mu \neq \mu_0$. We can use t statistic.
- **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**:

© SUZUKI, Atsushi

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) and determine which statistic to use. For example, if we are interested in the true expectation, the null hypothesis is $\mu = \mu_0$, where μ is the unknown true expectation and μ_0 is a value, which we decide. The alternative hypothesis is $\mu \neq \mu_0$. We can use t statistic.
- **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**.

© SUZUKI, Atsushi

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 $\alpha = 0.05$.

© SUZUKI, Atsushi

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.

© SUZUKI, Atsushi

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

5 Statistical Test

- Failure of statistical test

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

5 **Statistical Test**

- Exercises

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.

© SUZUKI, Atsushi

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.

© SUZUKI, Atsushi

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.