2023.12.15

短期集中研修
『Rデータ解析自由自在（入門編）』

④確率論の基礎

三村　喬生

統計数理研究所
医療健康データ科学研究センター

昨日の内容

- データ科学とは
- データ科学のツール
- Rを始めよう

今日の内容

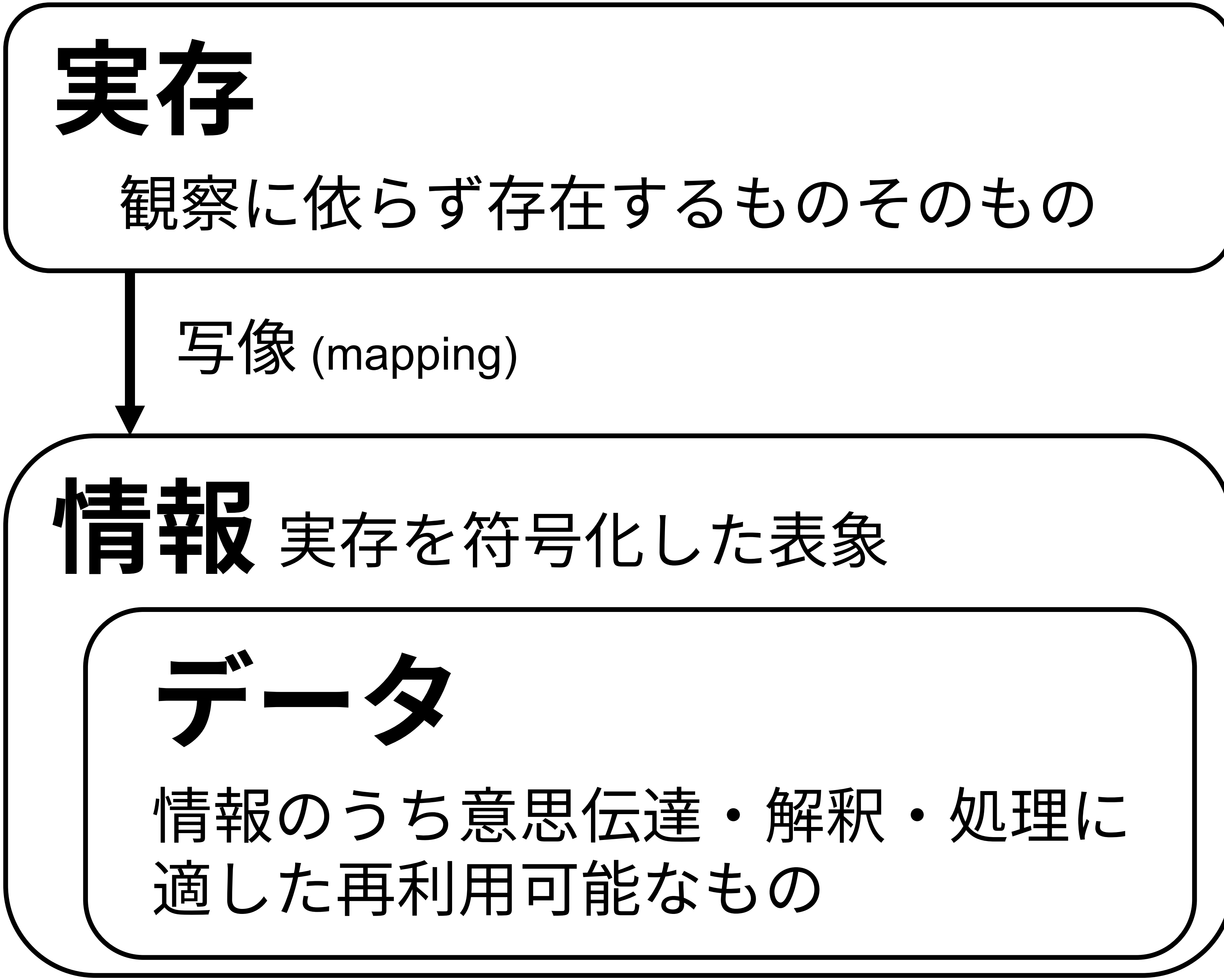
- 確率の話
- 回帰モデルの話
- 分散分析の話

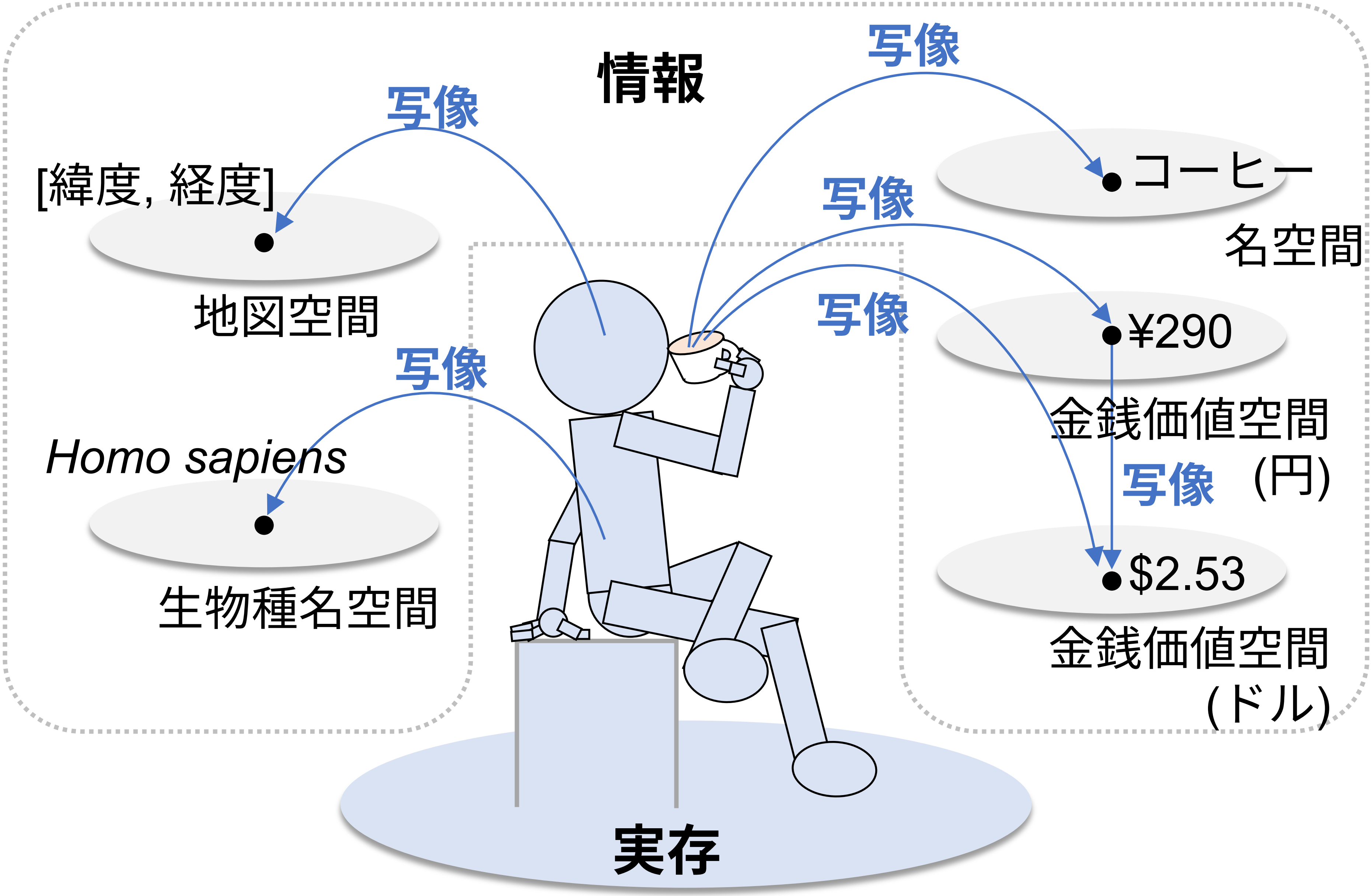
データ科学

対象に内在する構造を体系的に切り分け
共有可能な知識に変換すること、その方法

（ある視点から）同じものを同じとし、
違うものを違うものとして分類整理する

情報（実存の写像）のうち意思伝達・解釈・
処理に適した再利用可能なもの

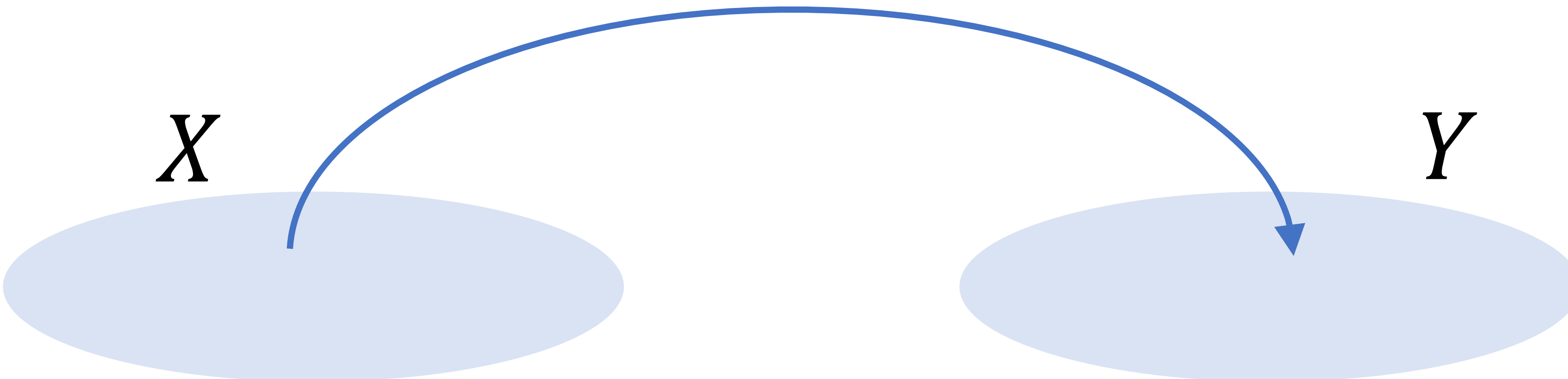

実存
観察に依らず存在するものそのもの
写像 (mapping)
情報
実存を符号化した表象
データ
情報のうち意思伝達・解釈・処理に適した再利用可能なもの



【写像】
ある集合の要素を他の集合のただ1つの要素に対応づける規則

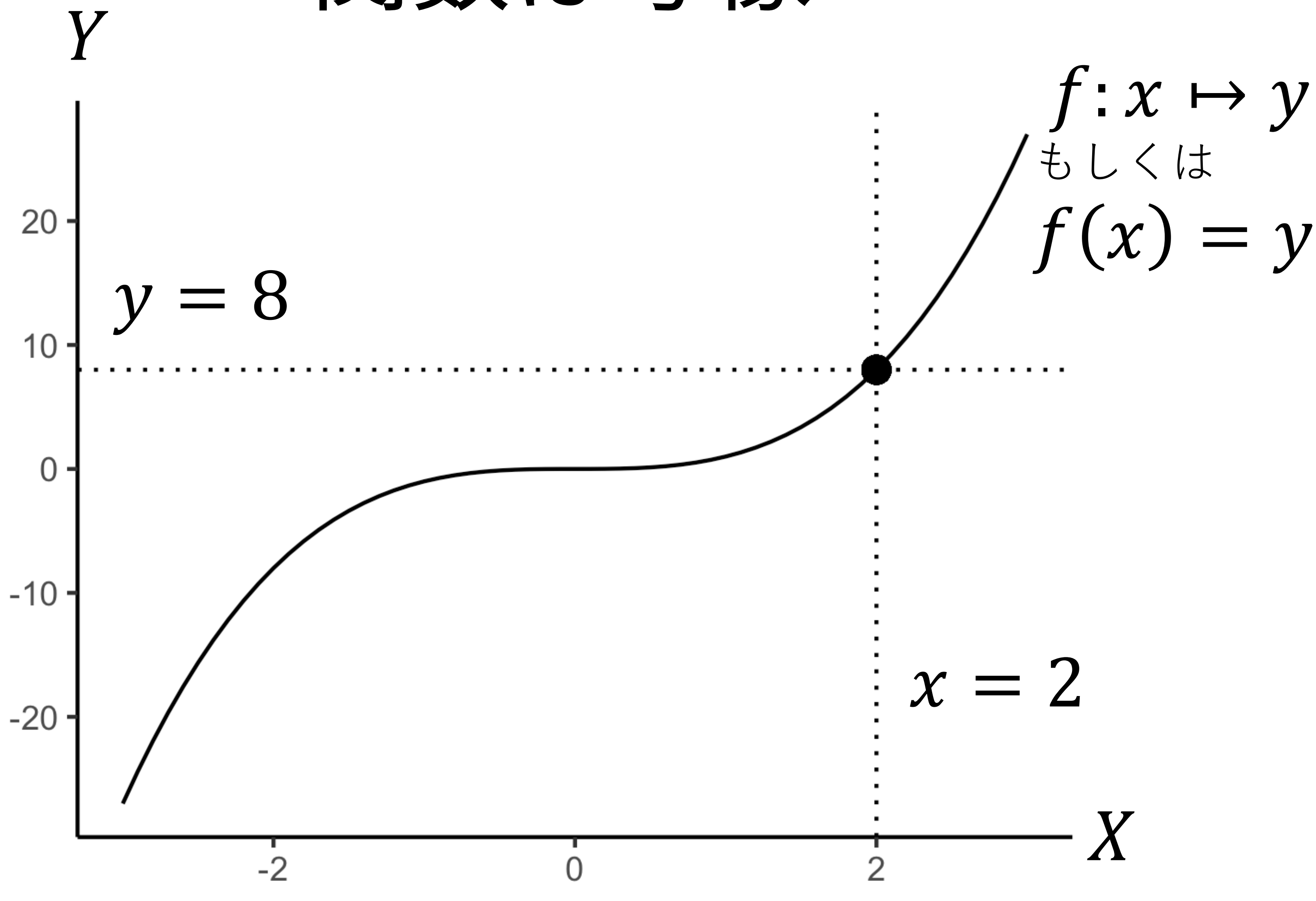
写像 mapping

$$f: X \to Y$$

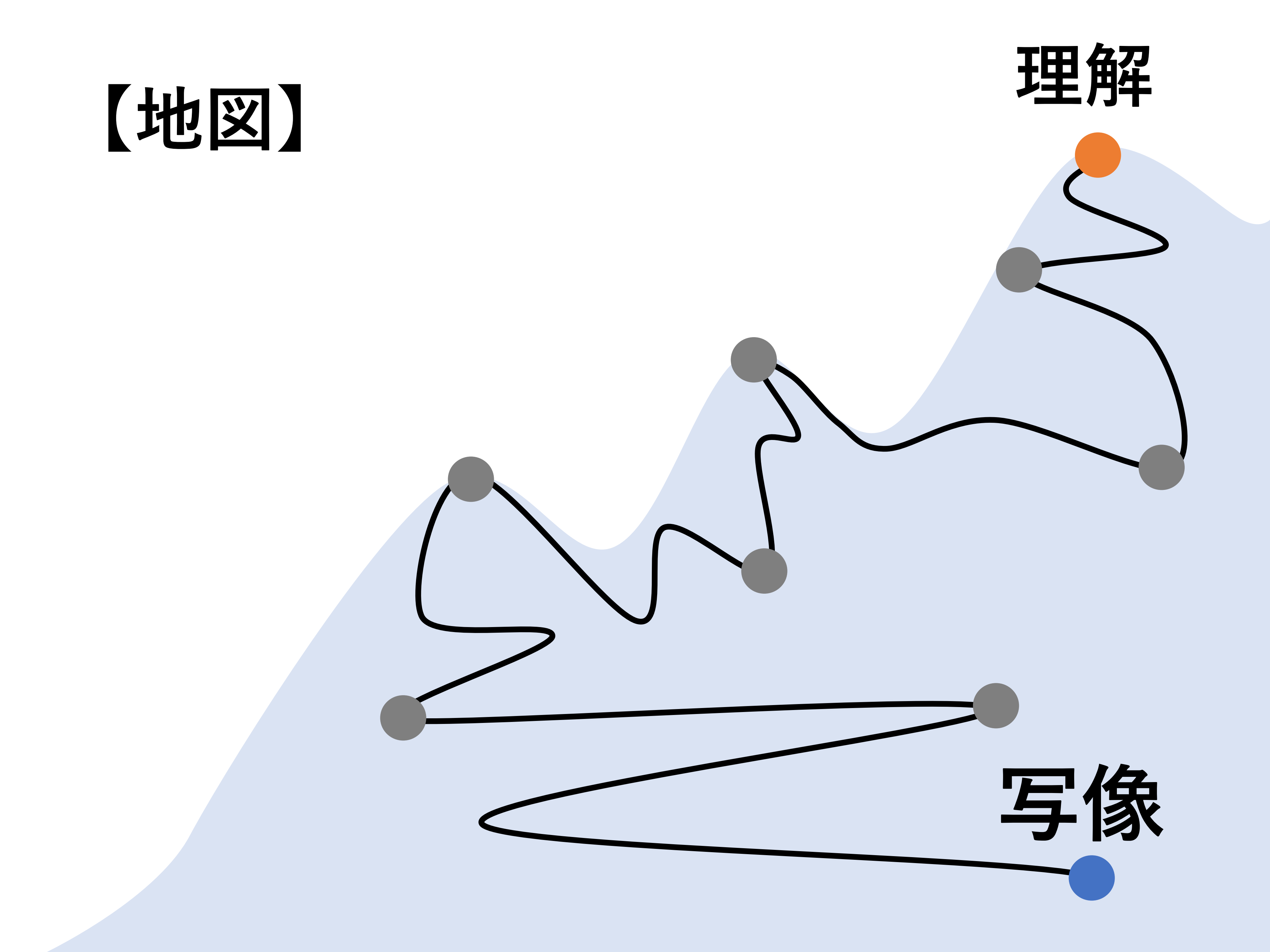


ある情報の集合の要素を、別の情報の集合の
ただ1つの要素に対応づけるプロセス

関数は写像



【写像】
ある集合の要素を他の集合のただ1つの要素に対応づける規則


【地図】
理解
写像

【目標】

$$X \sim N(\mu = 0, \sigma = 1)$$

「**確率変数**X は**正規分布**$N(0,1)$に従う」

を完全に理解しよう。

よく見る式（正規分布といえばコレ）

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

よく見る式（正規分布といえばコレ）

平均μ

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

標準偏差σ

よく見る式（正規分布といえばコレ）

平均μ

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

標準偏差σ

あなたはだあれ？

よく見る式（正規分布といえばコレ）

平均μ

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

標準偏差σ

あなたはだあれ？

やってみよう

Rでこの式の挙動を調べてグラフを描画する

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

```r
## packages ----
library(tidyverse)

## parameteres ----
mu <- 0
sd <- 1

## data ----
df <-
  tibble(
    x = seq(-3, 3, by = 0.01)
    px = ***
  )
```

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

```r
## packages ----
library(tidyverse)

## parameteres ----
mu <- 0
sd <- 1

## data ----
df <-
  tibble(
    x = seq(-3, 3, by = 0.01)
    px = exp(-(x - mu)^2 / 2 * sd^2) /
          sqrt(2 * pi * sd^2)
  )
```

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

```
## packages ----
library(tidyverse)

## parameteres ----
mu <- 0
sd <- 1

## function ----
Norm_p <- function(x, mu, sd){
  ****
}
```

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

```r
## packages ----
library(tidyverse)

## parameteres ----
mu <- 0
sd <- 1

## function ----
Norm_p <- function(x, mu, sd){
  exp(-(x - mu)^2 / 2 * sd^2) / sqrt(2 * pi * sd^2)
}

## test ----
seq(0, 5, by = 1) %>% Norm_p() %>% round(digits = 2)
#> [1] 0.40 0.24 0.05 0.00 0.00 0.00
```

```r
## packages ----
library(tidyverse)

## parameteres ----
mu <- 0
sd <- 1

## function ----
Norm_p <- function(x, mu, sd){
  exp(-(x - mu)^2 / 2 * sd^2) / sqrt(2 * pi * sd^2)
}

## data ----
df <-
  tibble(
    x = seq(-5, 5, by = 0.01),
    px = Norm_p(x = x, mu = mu, sd = sd)
  )
```

```r
df <-
  data.frame(
    x = seq(-5, 5, by = 0.01)
  ) %>%
  mutate(px = Norm_p(x = x, mu = mu, sd = sd))
```

```
## packages ----
library(tidyverse)

## parameteres ----
mu <- 0
sd <- 1

## function ----
Norm_p <- function(x, mu, sd){
  exp(-(x - mu)^2 / 2 * sd^2) / sqrt(2 * pi * sd^2)
}

## data ----
df <-
  tibble(
    x = seq(-5, 5, by = 0.01),
    px = Norm_p(x = x, mu = mu, sd = sd)
  )
```

tibbleに存在しない変数なので
環境中のオブジェクトを探す

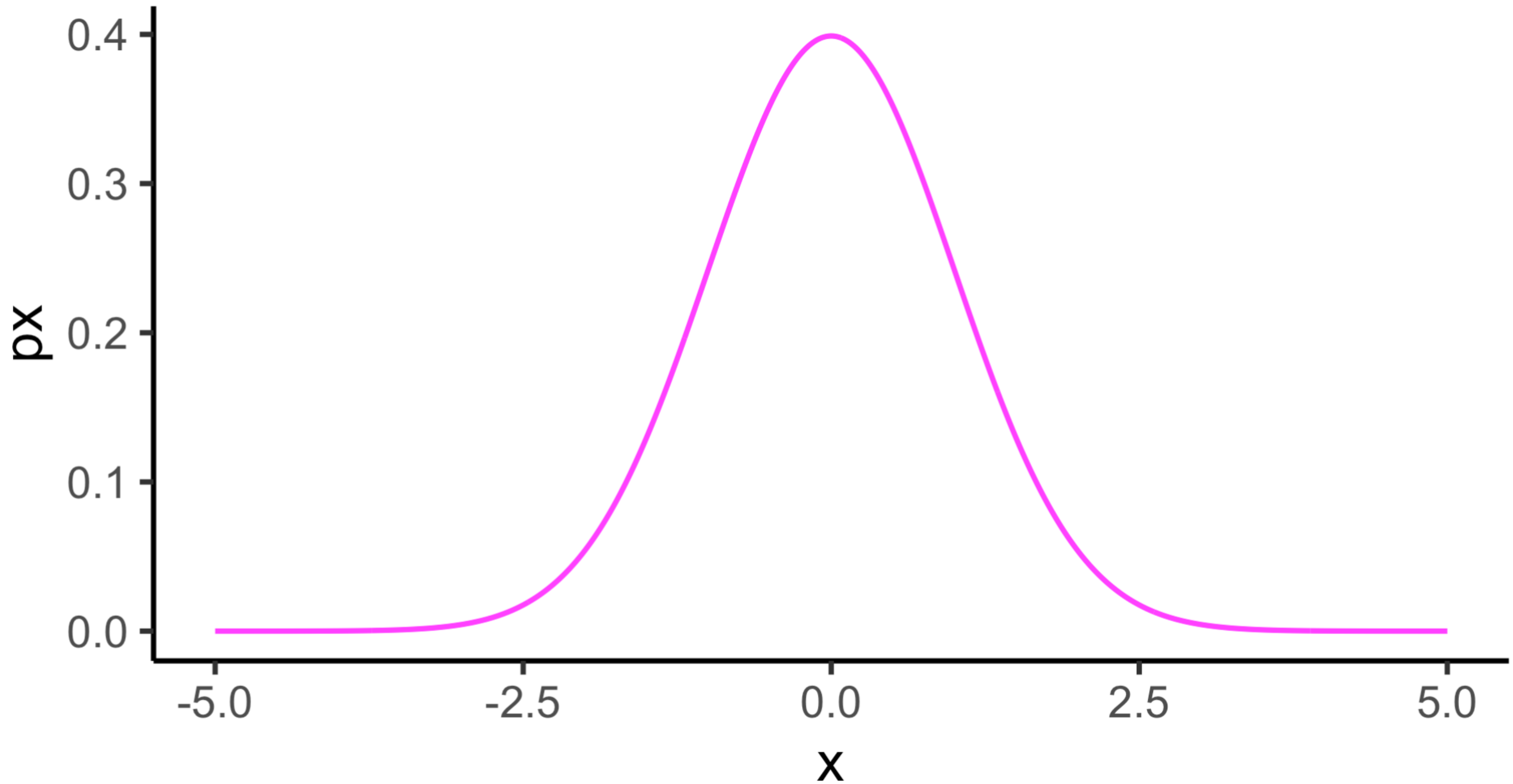
遅延評価 & 非標準評価

```r
## packages ----
library(tidyverse)

## parameteres ----
mu <- 0
sd <- 1

## function ----
Norm_p <- function(x, mu, sd){
  exp(-(x - mu)^2 / 2 * sd^2) / sqrt(2 * pi * sd^2)
}

## data ----
df <-
  tibble(
    x = seq(-5, 5, by = 0.01),
    px = Norm_p(x = x, mu = mu, sd = sd)
  )

## visualization ----
ggplot(data = df) +
  aes(x = x, y = px) +
  geom_path(color = "magenta") +
  theme_classic()
```

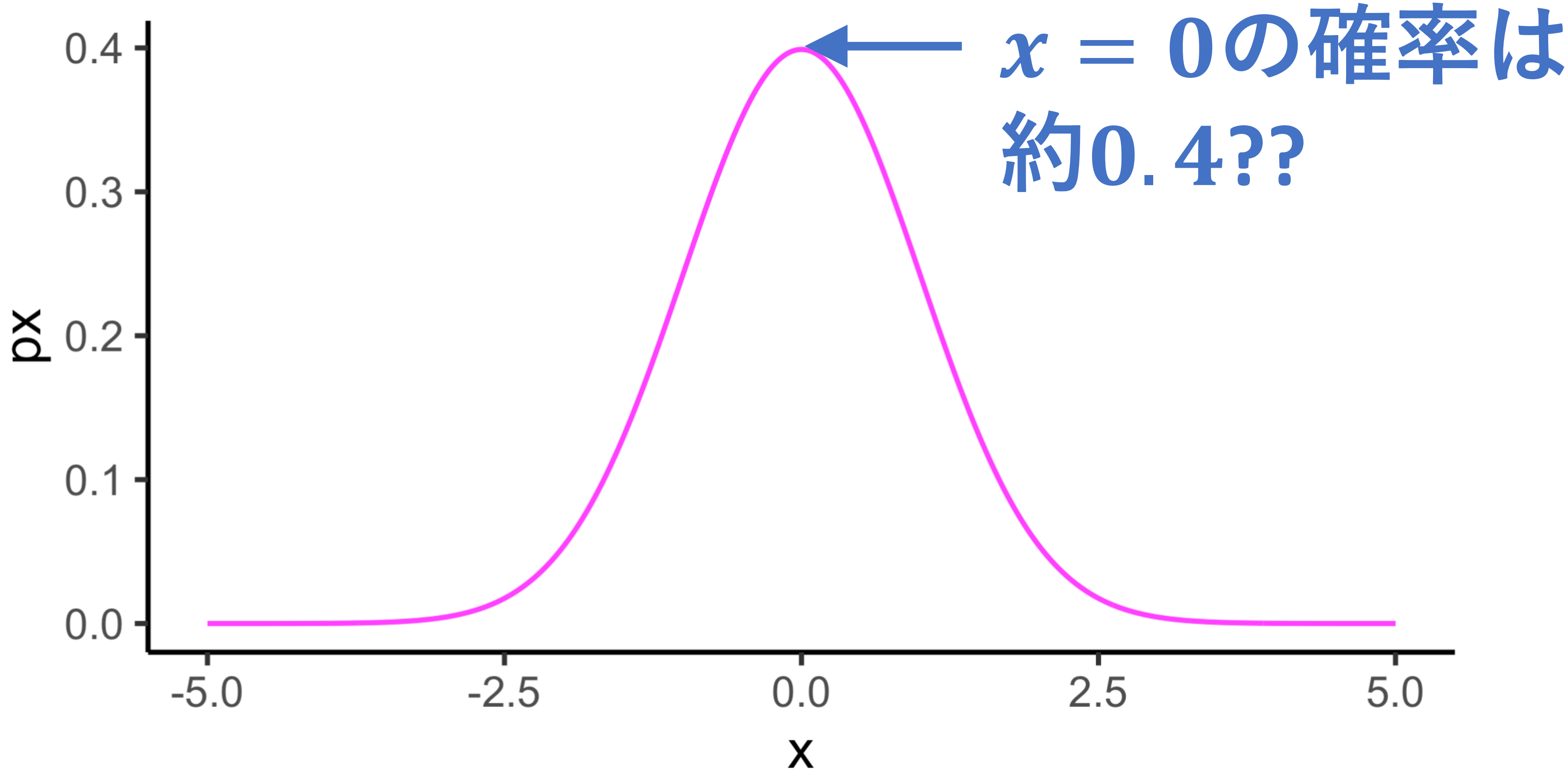
tibbleに存在しない変数なので
環境中のオブジェクトを探す

遅延評価 & 非標準評価

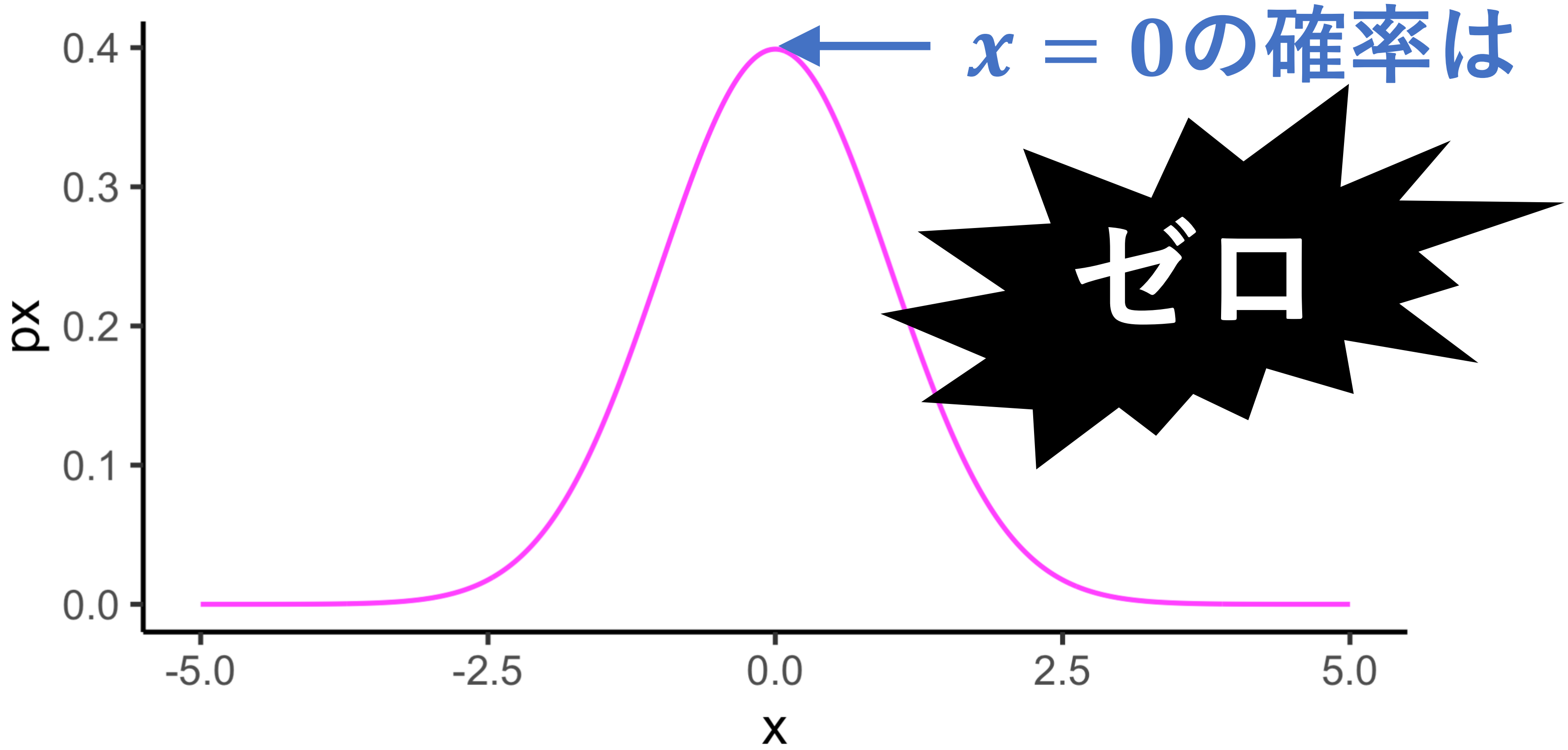
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$



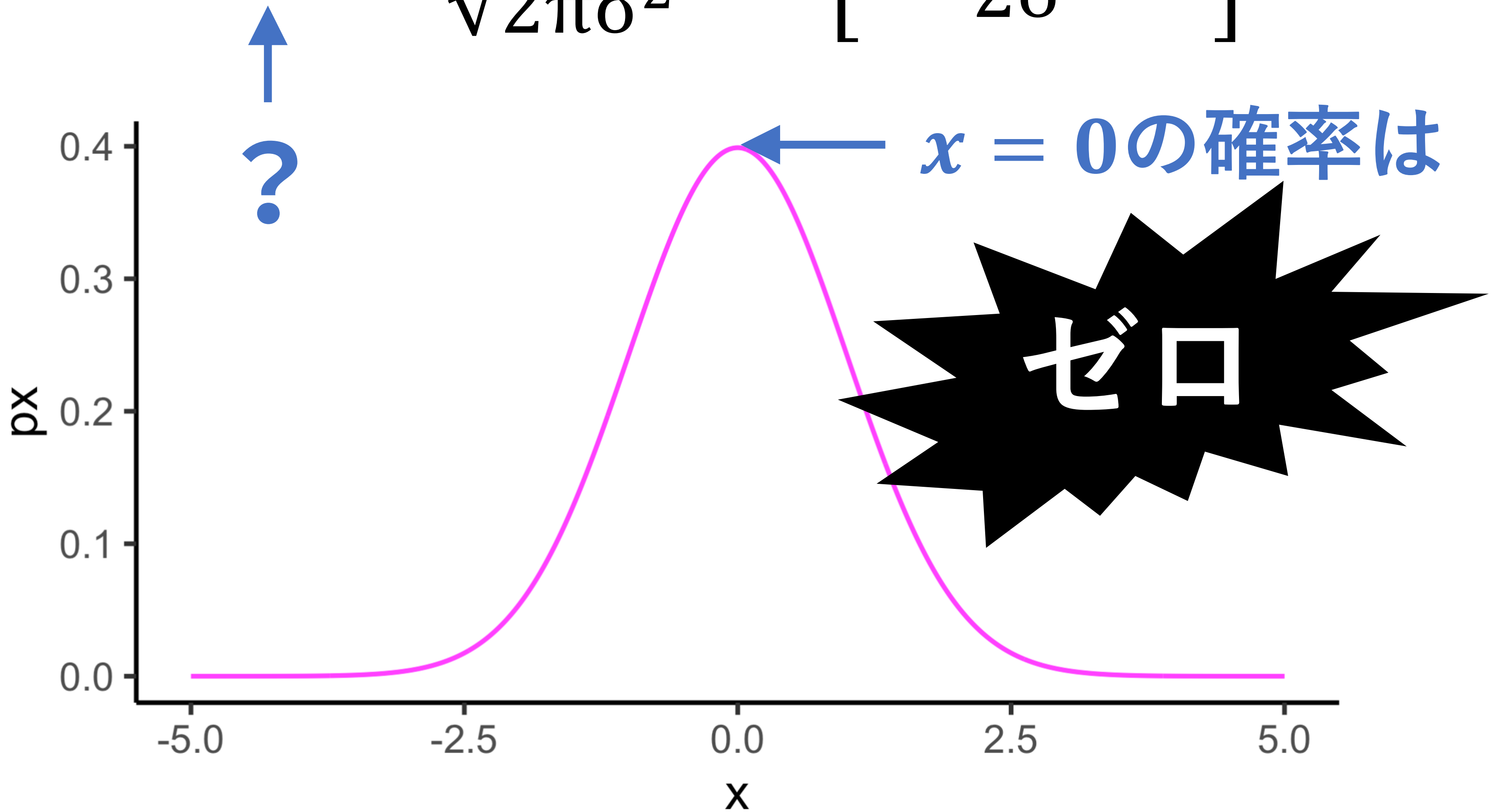
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$



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

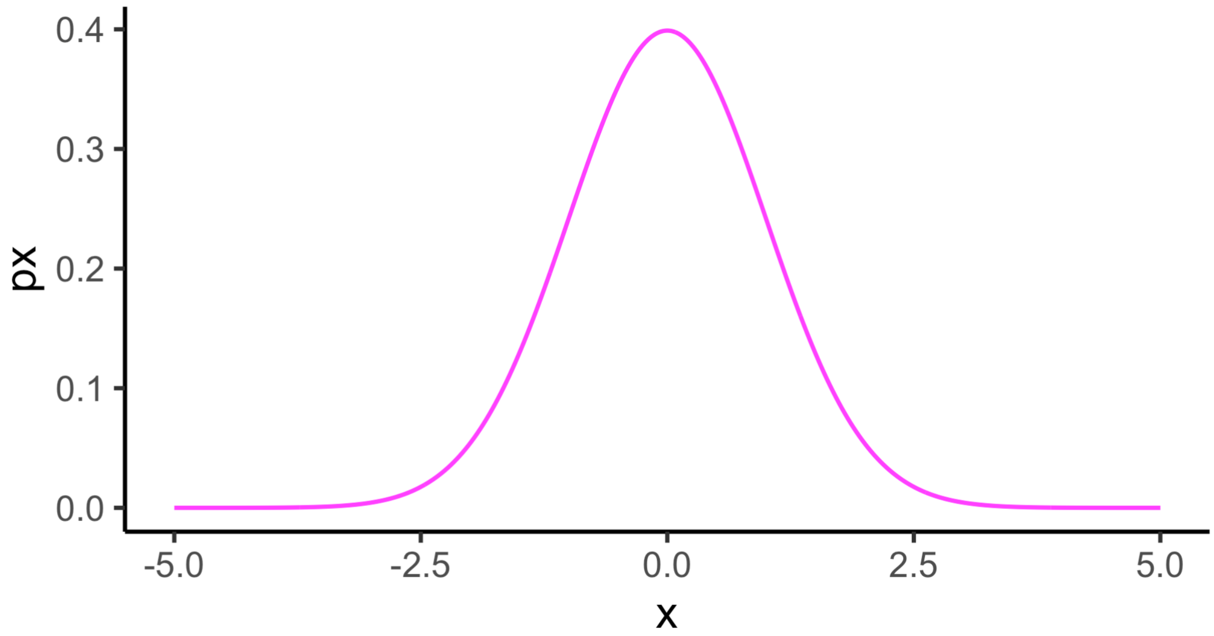


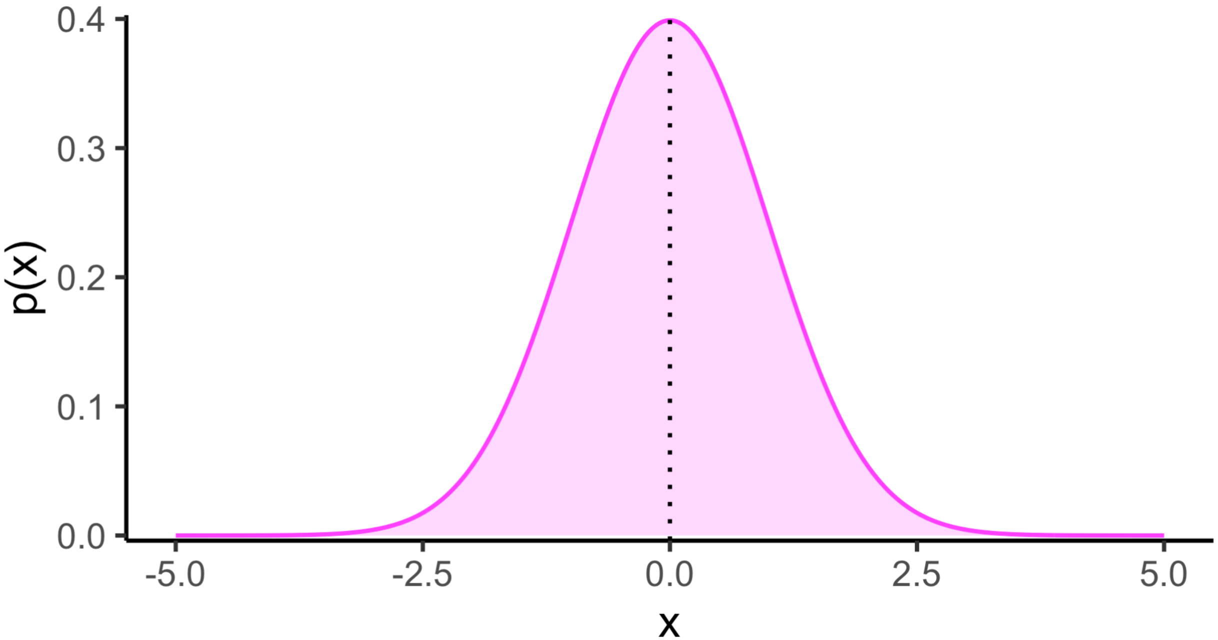
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

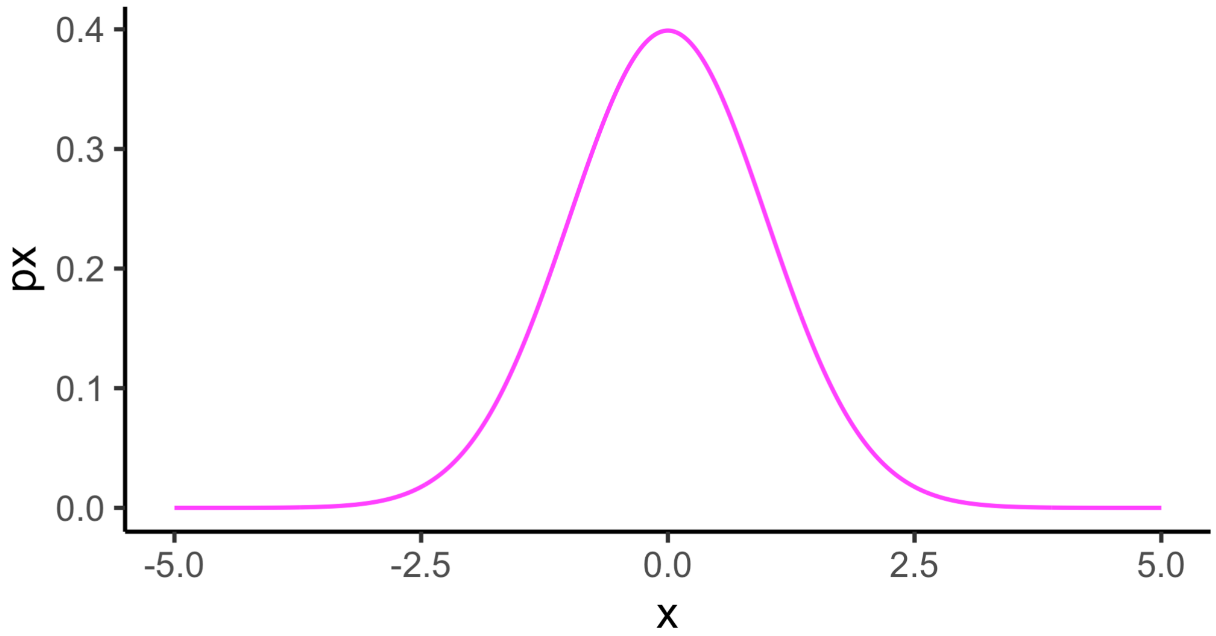


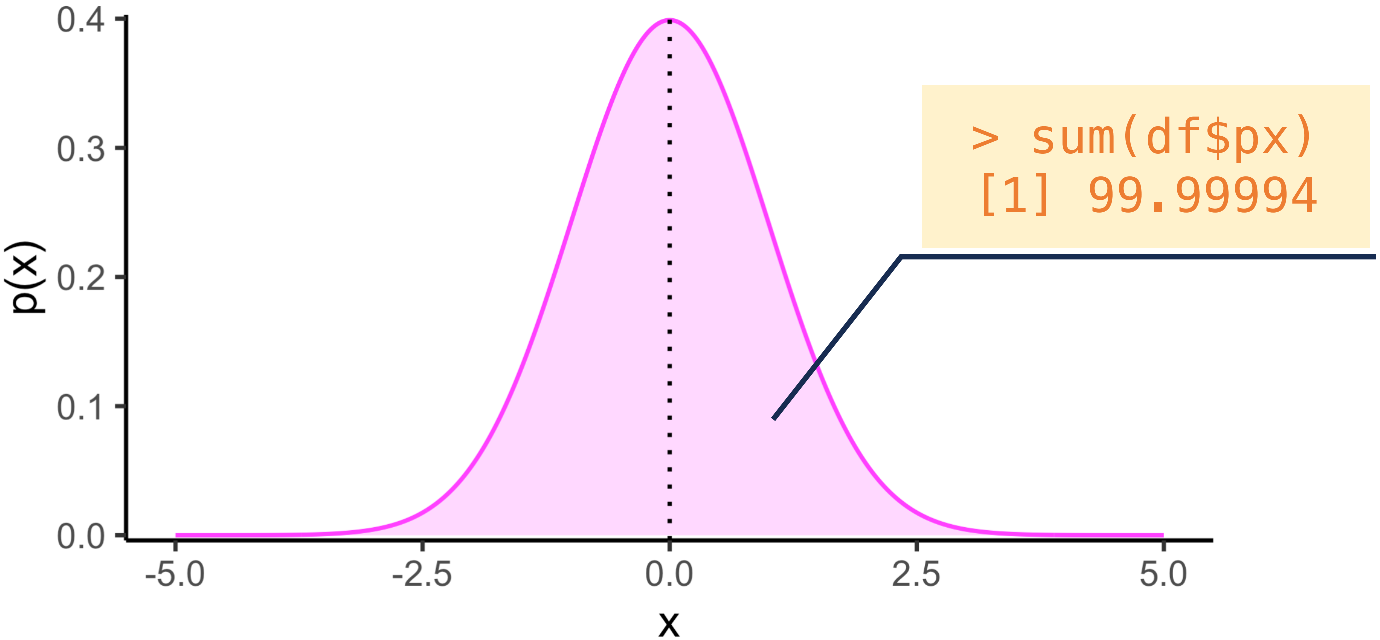
```r
## visualization ----
ggplot(data = df) +
  aes(x = x, y = px) +
  geom_path(color = "magenta") +
  theme_classic()
```

```r
## visualization ----
ggplot(data = df) +
  aes(x = x, y = px) +
  geom_ribbon(
    mapping = aes(ymin = 0, ymax = px),
    fill = "magenta", alpha = 0.2
  ) +
  geom_path(color = "magenta") +
  geom_vline(xintercept = 0,
             linetype = "dotted") +
  scale_y_continuous(limits = c(0, NA),
                     expand = c(0.01, 0)) +
  theme_classic() +
  labs(y = "p(x)")
```


0.4
0.3
0.2
0.1
0.0
px
-5.0
-2.5
0.0
2.5
5.0
x


0.4
0.3
0.2
0.1
0.0
p(x)
-5.0
-2.5
0.0
2.5
5.0
x


0.4
0.3
0.2
0.1
0.0
px
-5.0
-2.5
0.0
2.5
5.0
x

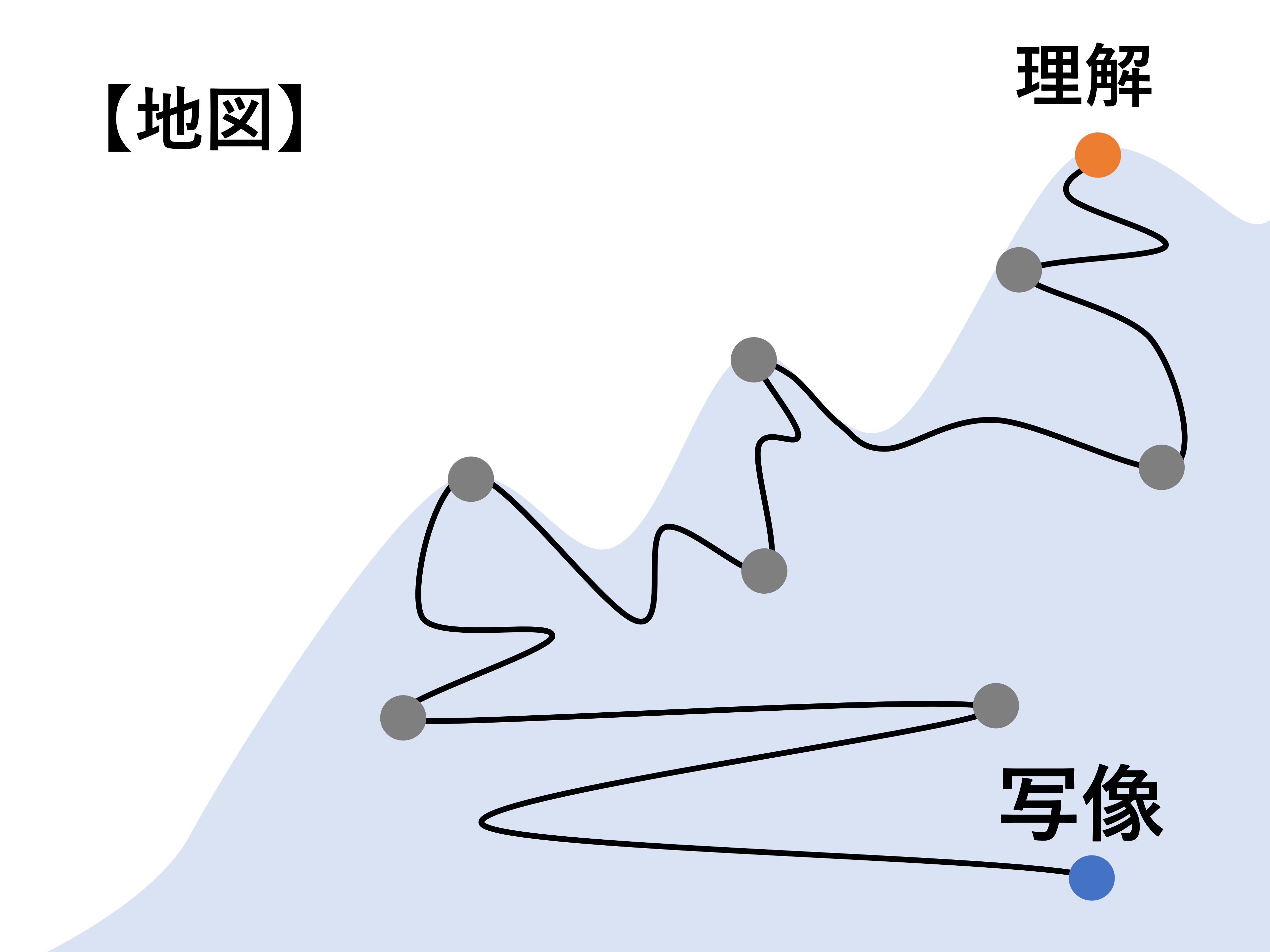

0.4
0.3
0.2
0.1
0.0
p(x)
-5.0
-2.5
0.0
2.5
5.0
x
> sum(df$px)
[1] 99.99994

【目標】

$$X \sim N(\mu = 0, \sigma = 1)$$

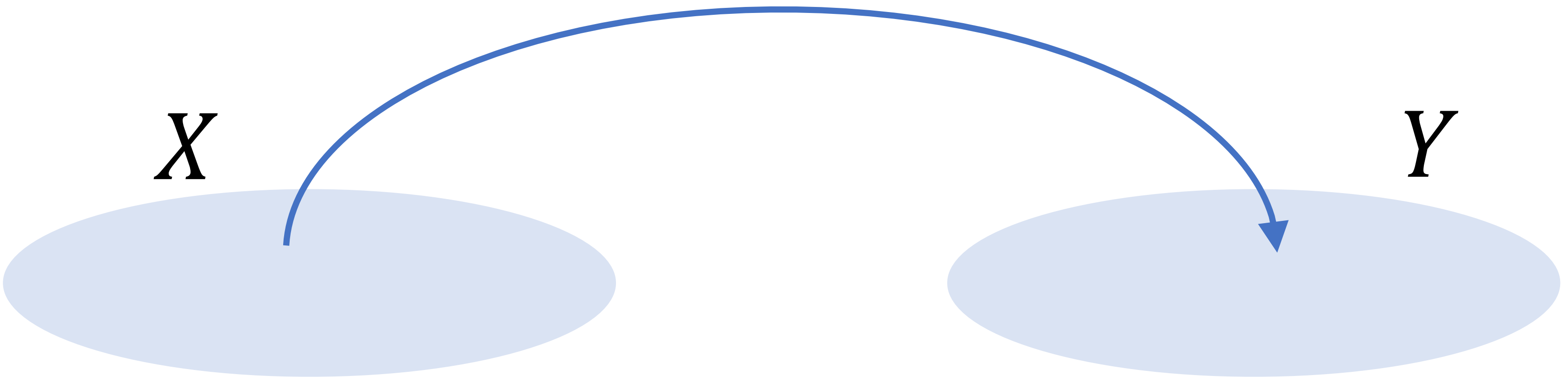
「**確率変数**X は**正規分布**$N(0,1)$に従う」

を完全に理解しよう。


【地図】
理解
写像

写像 mapping

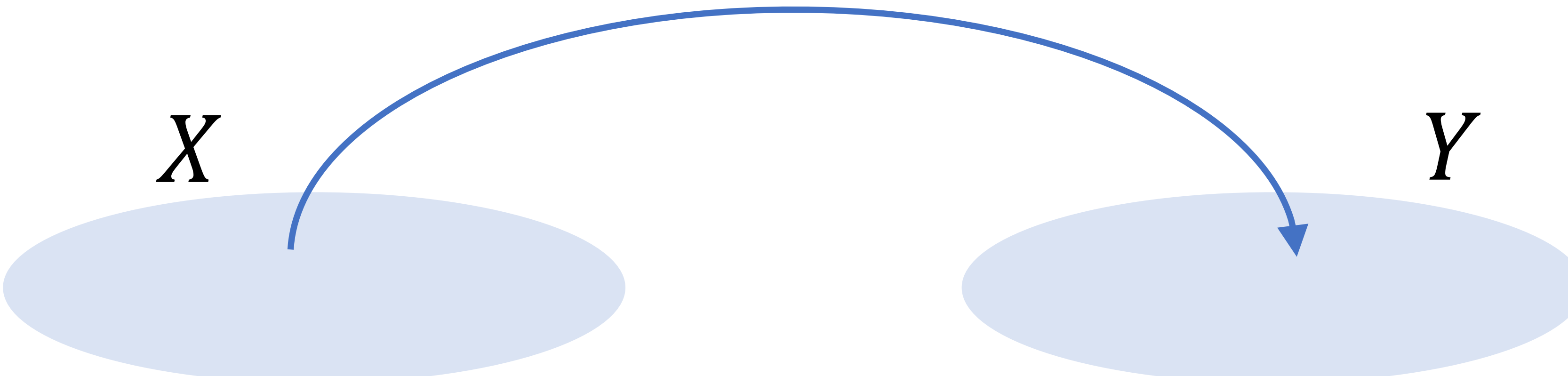
$$f: X \to Y$$



ある情報の集合の要素を、別の情報の集合の
ただ1つの要素に対応づけるプロセス

写像 mapping

$$f: X \to Y$$



domain
集合Xは写像fの**定義域**

final set
集合Yは写像fの**終域**

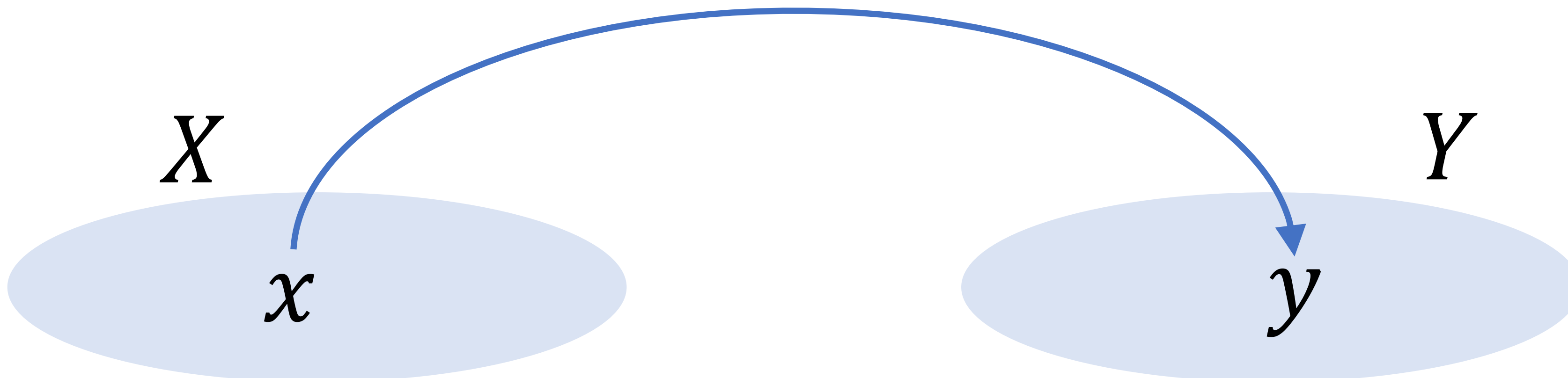
「写像fは集合Xで**定義**される」
「集合Xで**定義**される写像f」などという

写像 mapping

$$f : X \to Y$$

もしくは

$$f : x \mapsto y$$



element

xは集合Xの**要素**

$x \in X$

yは集合Yの**要素**

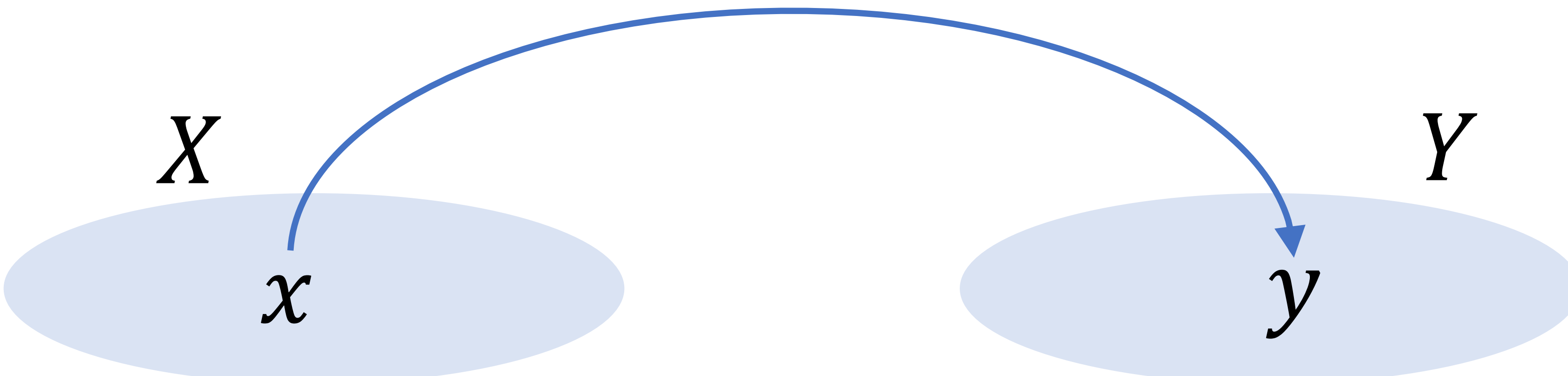
$y \in Y$

写像 mapping

$$f: X \to Y$$

もしくは

$$f: x \mapsto y$$



写像fによって定義域Xの要素xに対応づけられた終域Yの要素yを「fによるxの**像**y」と呼ぶ。
image

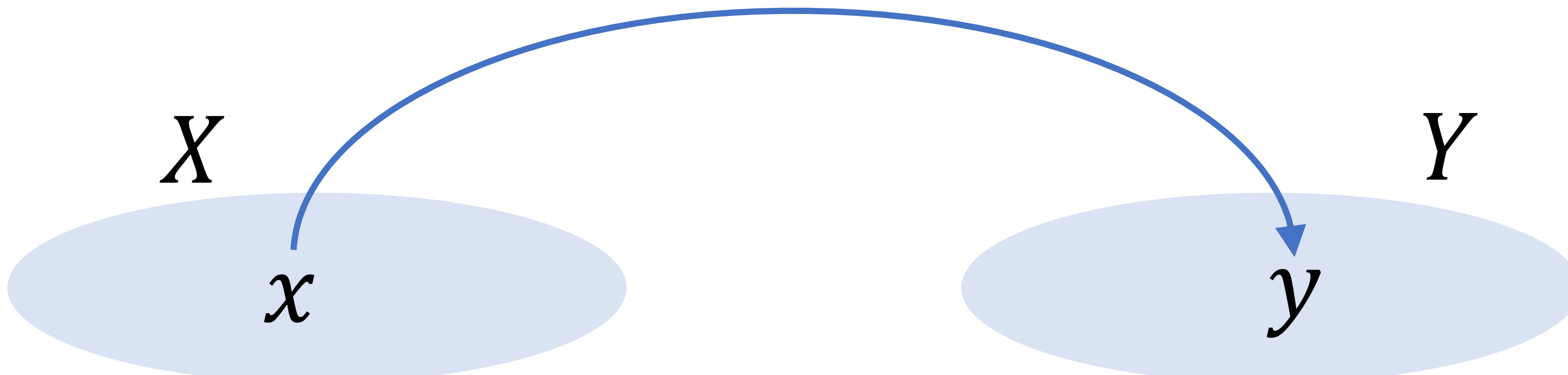
$$f(x) = y$$

写像 mapping

$$f: X \to Y$$

もしくは

$$f: x \mapsto y$$



写像fによって終域Yの要素yに対応づけられる定義域Xの要素xの集合を「yのfによる**逆像**」と呼ぶ。

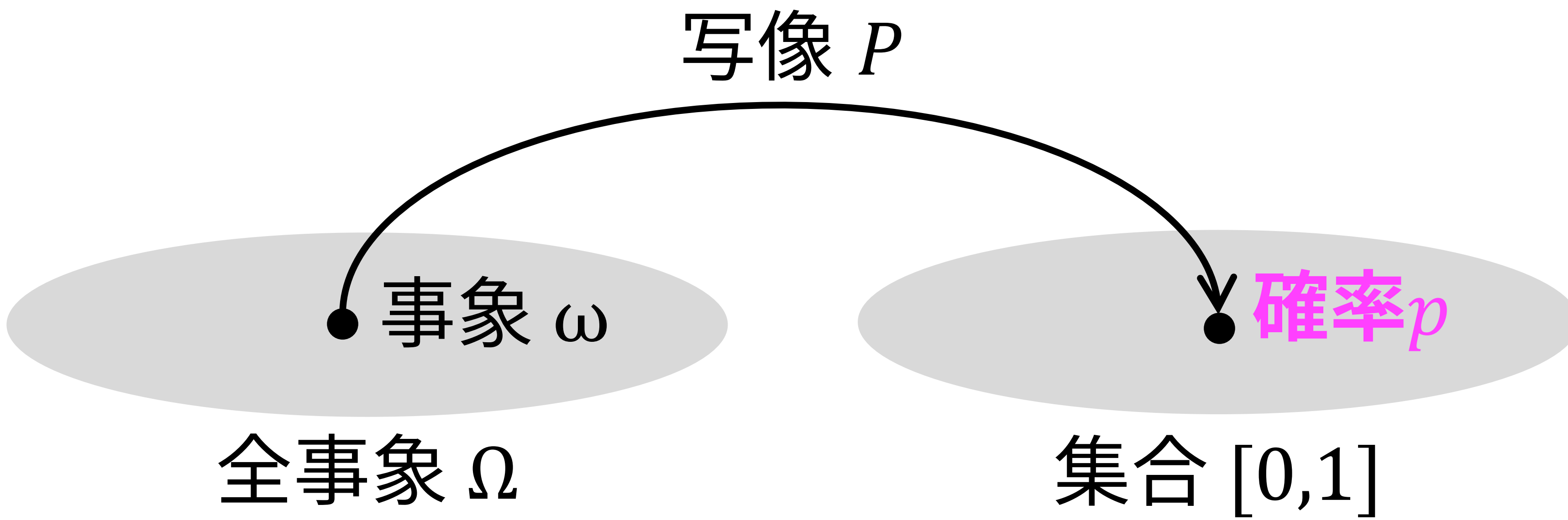
inverse image

$$f^{-1}(\mathrm{y}) = \{x \in X \mid f(x) = y\}$$

かくりつ　　　ぞう

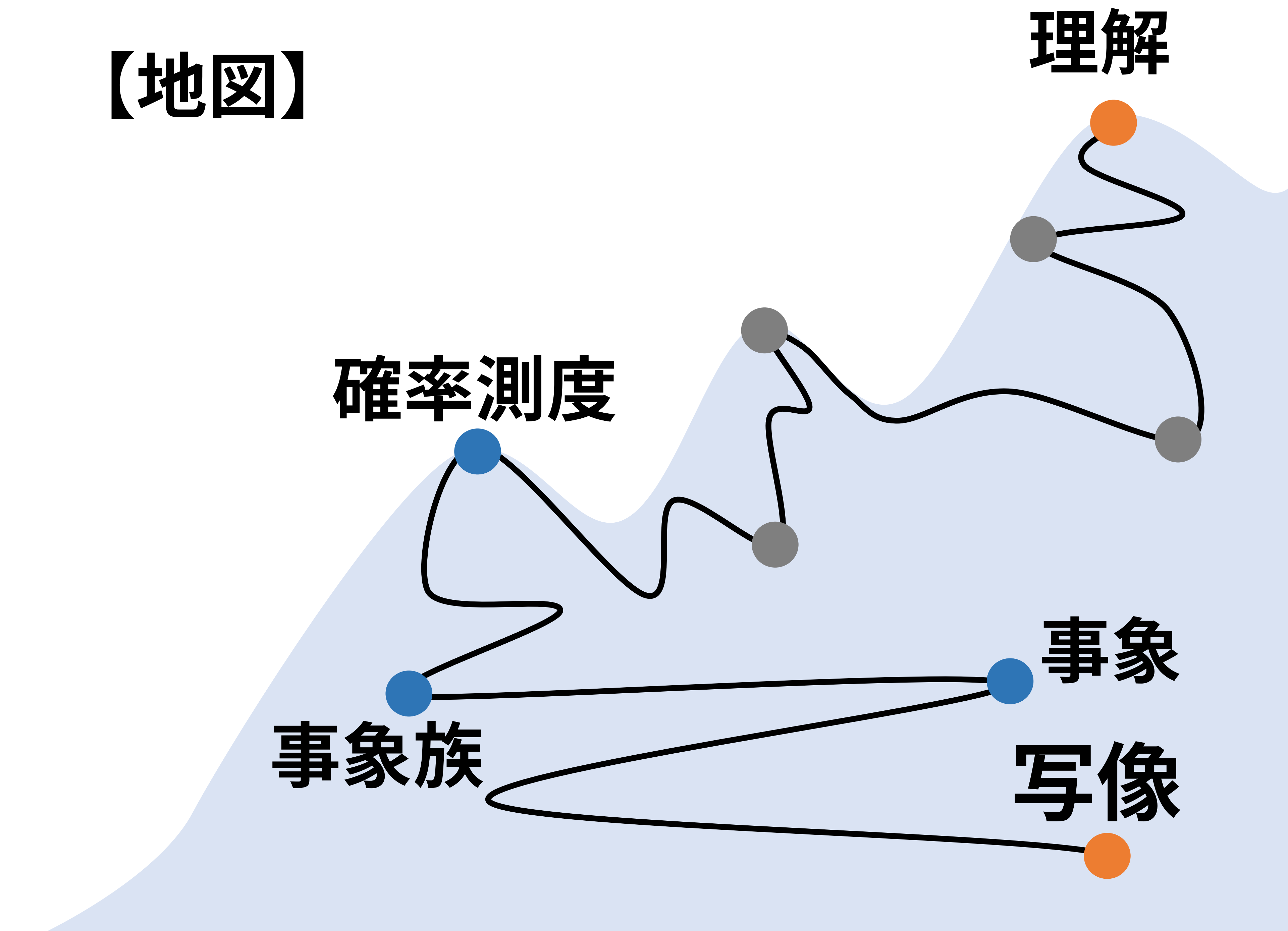
確率は像

確率は像

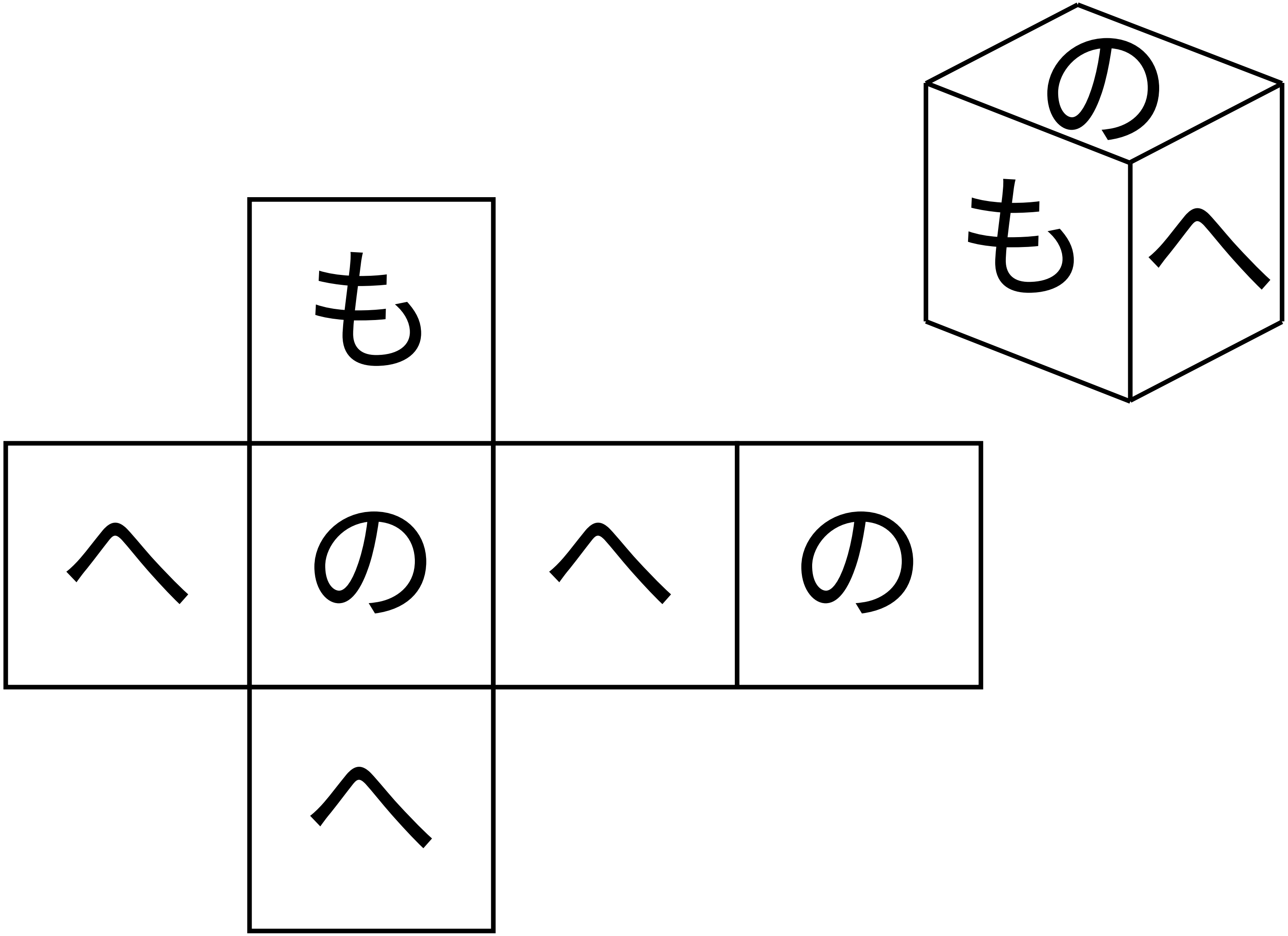


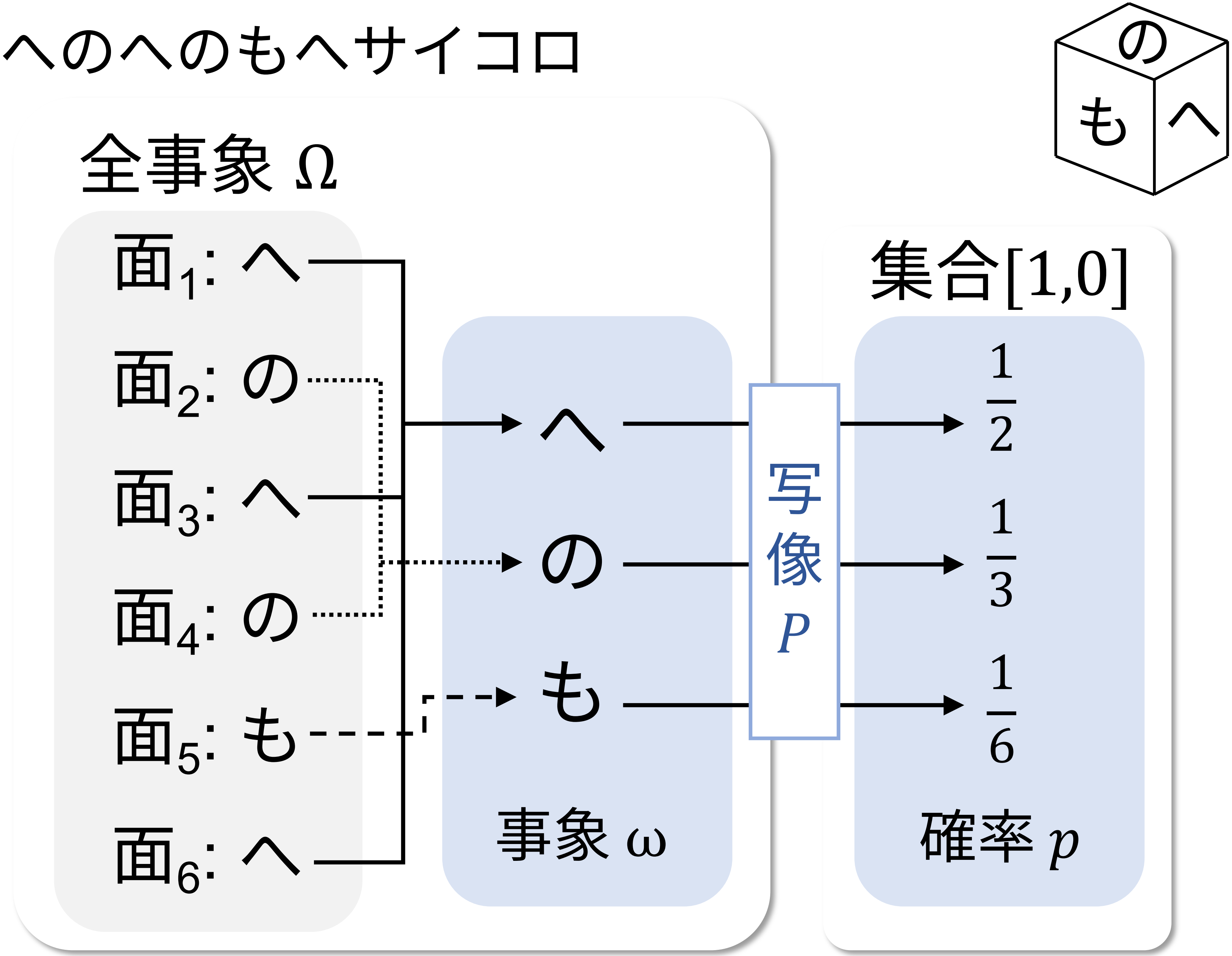
全事象Ωにおいて定義された写像Pによる、
事象$\omega \in \Omega$の像$p \in [0,1]$を「ωの**確率**p」と呼ぶ。

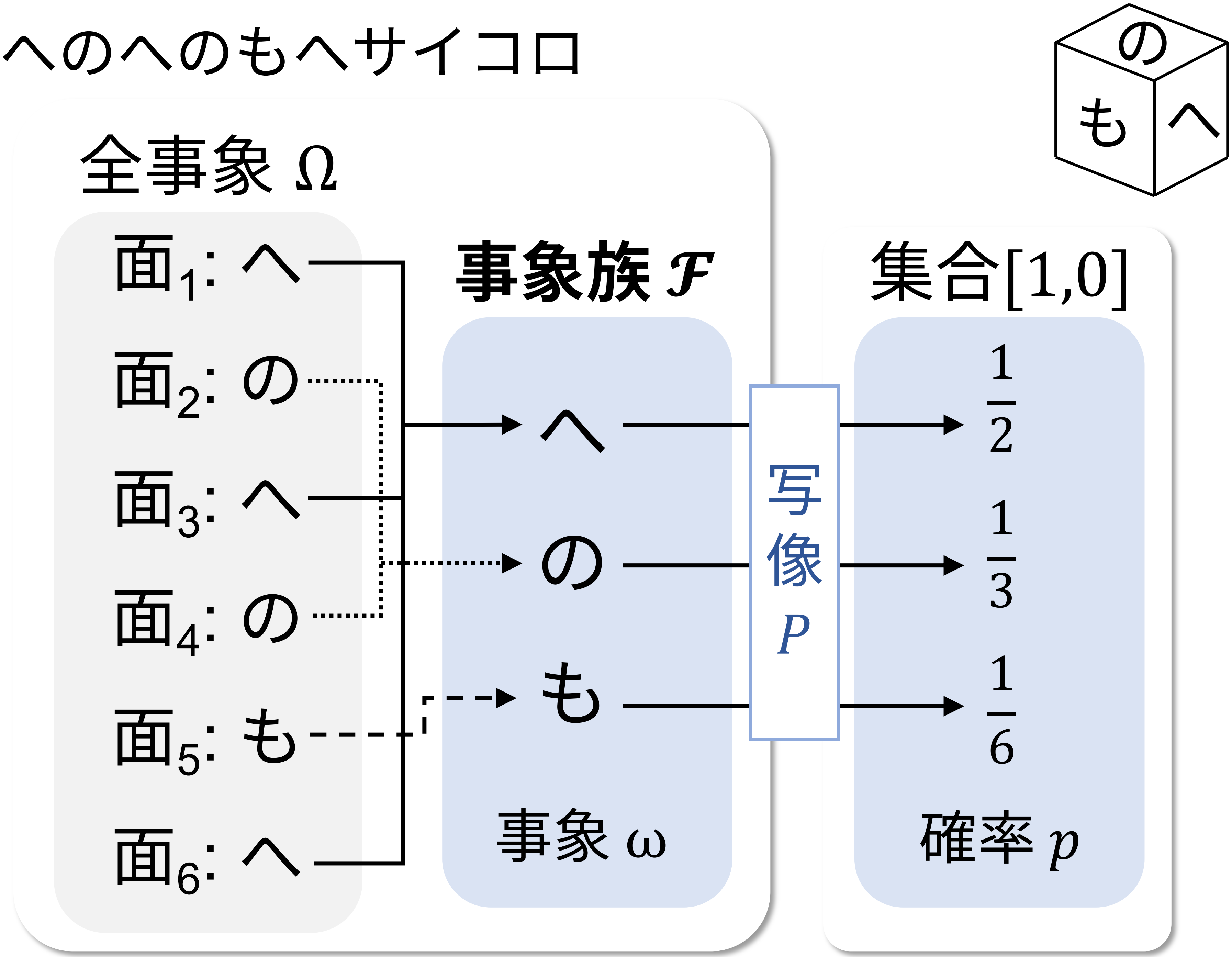
0以上1以下の実数の集合


【地図】
理解
確率測度
事象
事象族
写像

へのへのもへサイコロ

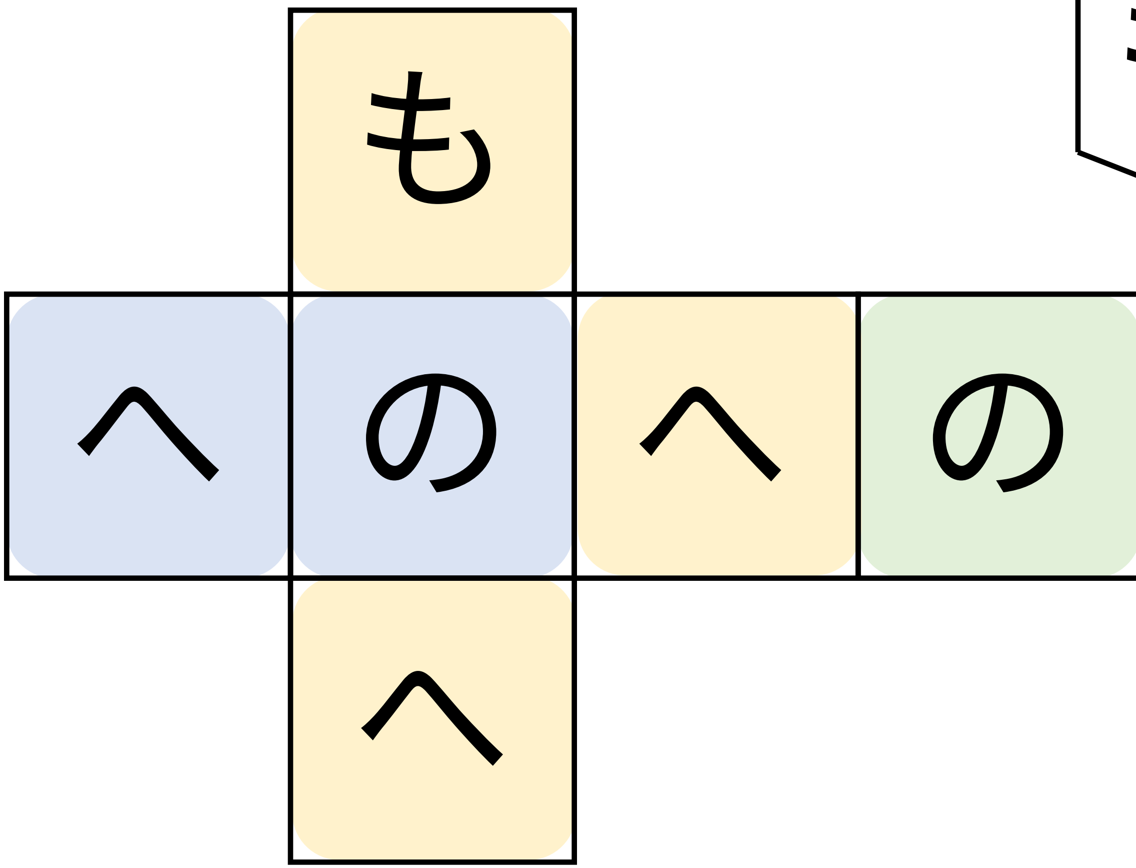


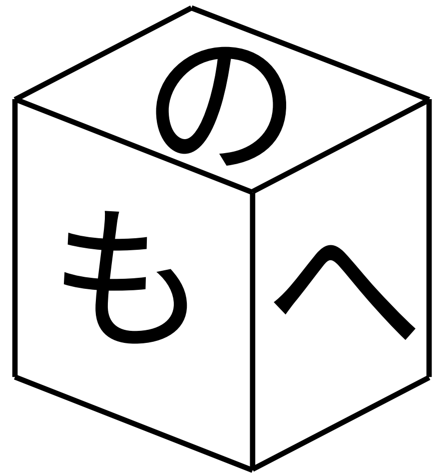

へのへのもへサイコロ
の
も
へ
全事象 Ω
面1: へ
面2: の
面3: へ
面4: の
面5: も
面6: へ
へ
の
も
事象 ω
写像 P
集合[1,0]
1/2
1/3
1/6
確率 p

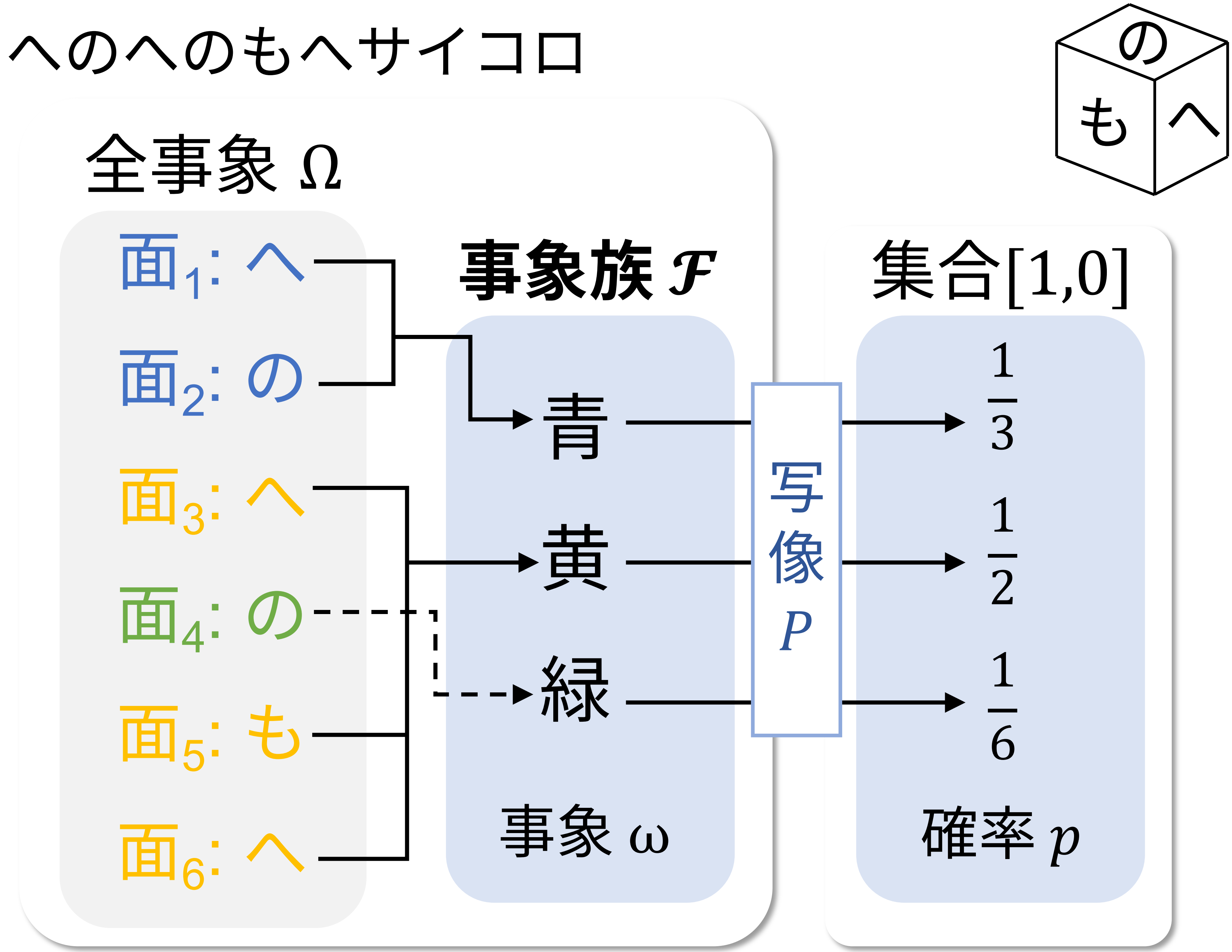


へのへのもへサイコロ
の
も
へ
全事象 Ω
面1: へ
面2: の
面3: へ
面4: の
面5: も
面6: へ
事象族 𝓕
へ
の
も
事象 ω
写像 P
集合[1,0]
1/2
1/3
1/6
確率 p

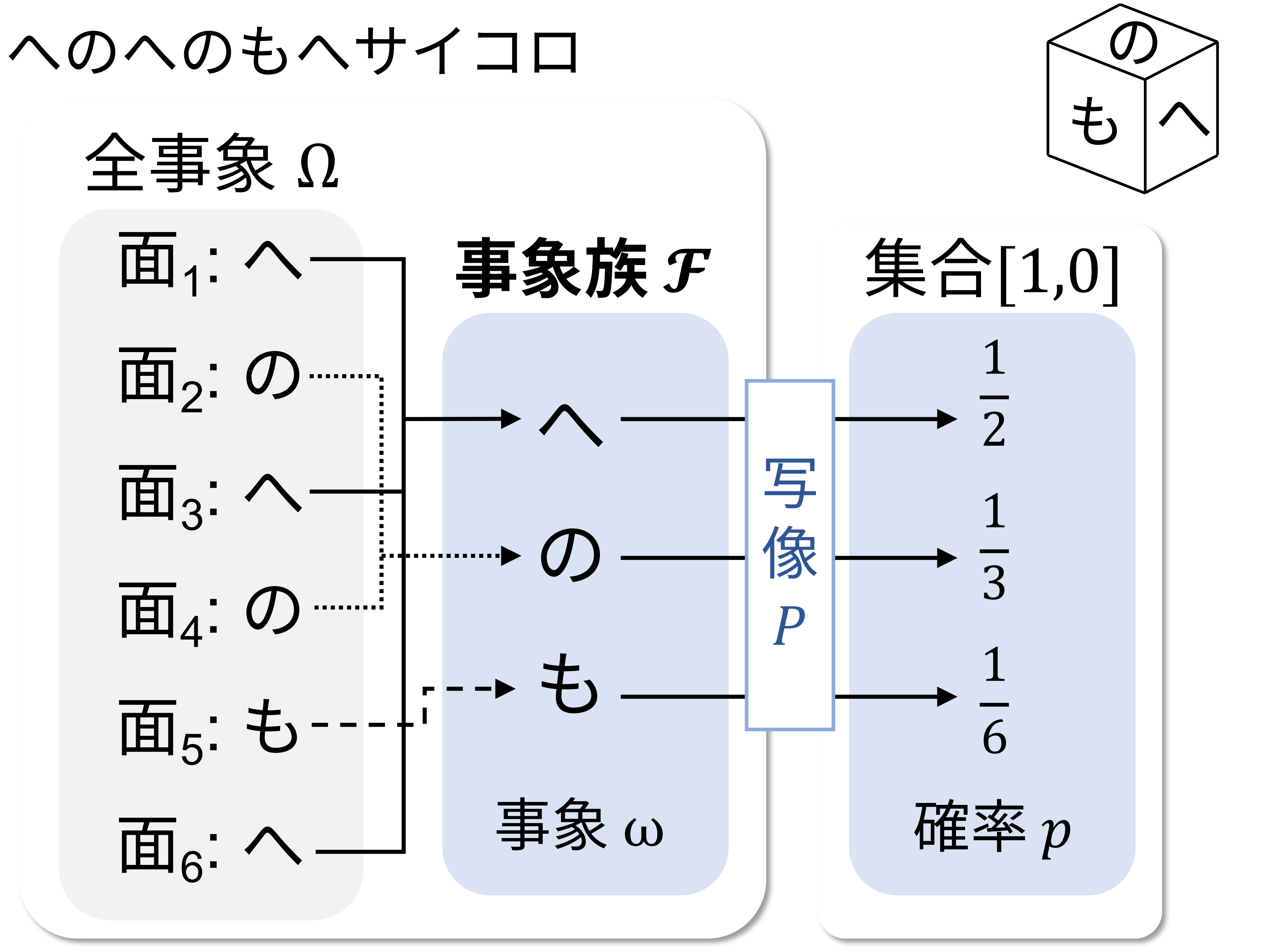
へのへのもへサイコロ



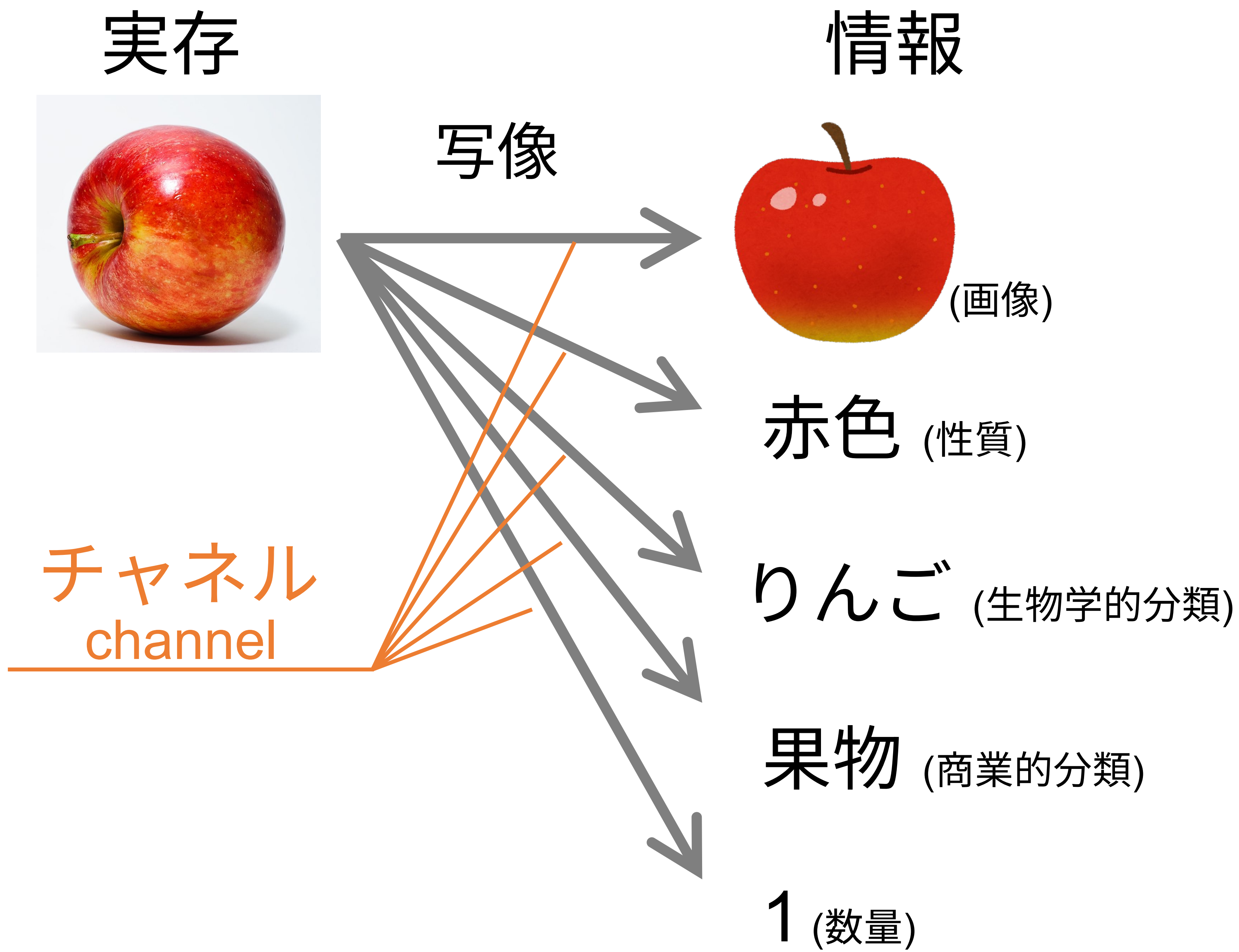




へのへのもへサイコロ
の
も
へ
全事象 Ω
面$_1$: へ
面$_2$: の
面$_3$: へ
面$_4$: の
面$_5$: も
面$_6$: へ
事象族 $\mathcal{F}$
青
黄
緑
事象 ω
写像 P
集合[1,0]
$\frac{1}{3}$
$\frac{1}{2}$
$\frac{1}{6}$
確率 p



へのへのもへサイコロ
の
も
へ
全事象 Ω
面1: へ
面2: の
面3: へ
面4: の
面5: も
面6: へ
事象族 𝓕
へ
の
も
事象 ω
写像 P
集合[1,0]
1/2
1/3
1/6
確率 p



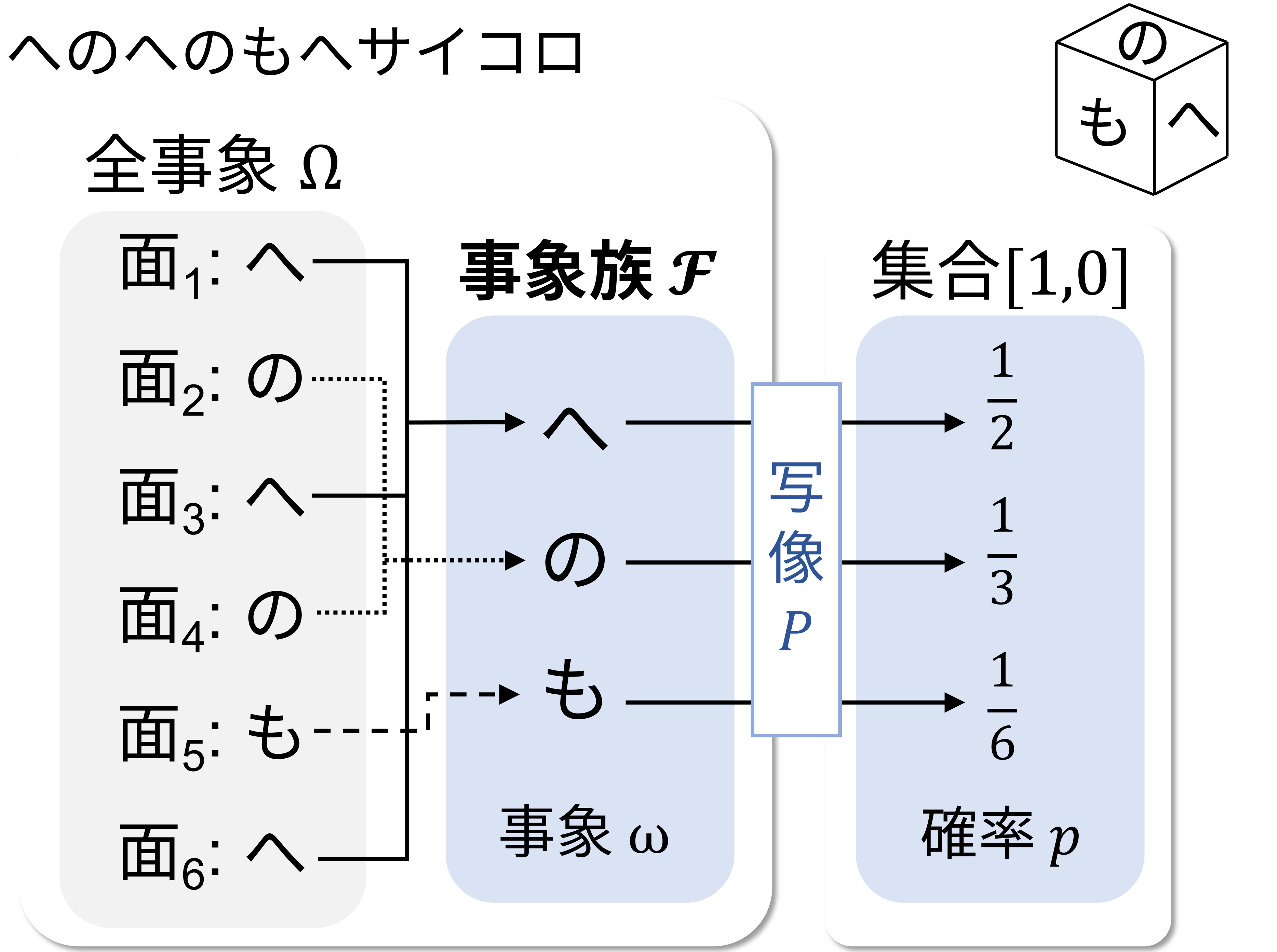
実存
情報
写像
(画像)
赤色 (性質)
チャネル
channel
りんご (生物学的分類)
果物 (商業的分類)
1 (数量)

事象族$\mathcal{F}$は写像Pのチャネル

全事象Ωのどの側面に対して事象 ωを割り当てるかは一般に一意には決まらない

「表に出た文字」という事象族 $\mathcal{F}$を割り振る場合、「表に出た色」という事象は全事象Ω に含まれるが事象族$\mathcal{F}$の定義において考慮されない。

確率 pは同一の全事象Ωに対して、事象族$\mathcal{F}$ごとに定義することができる。

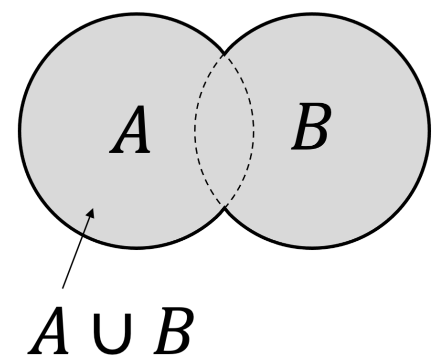


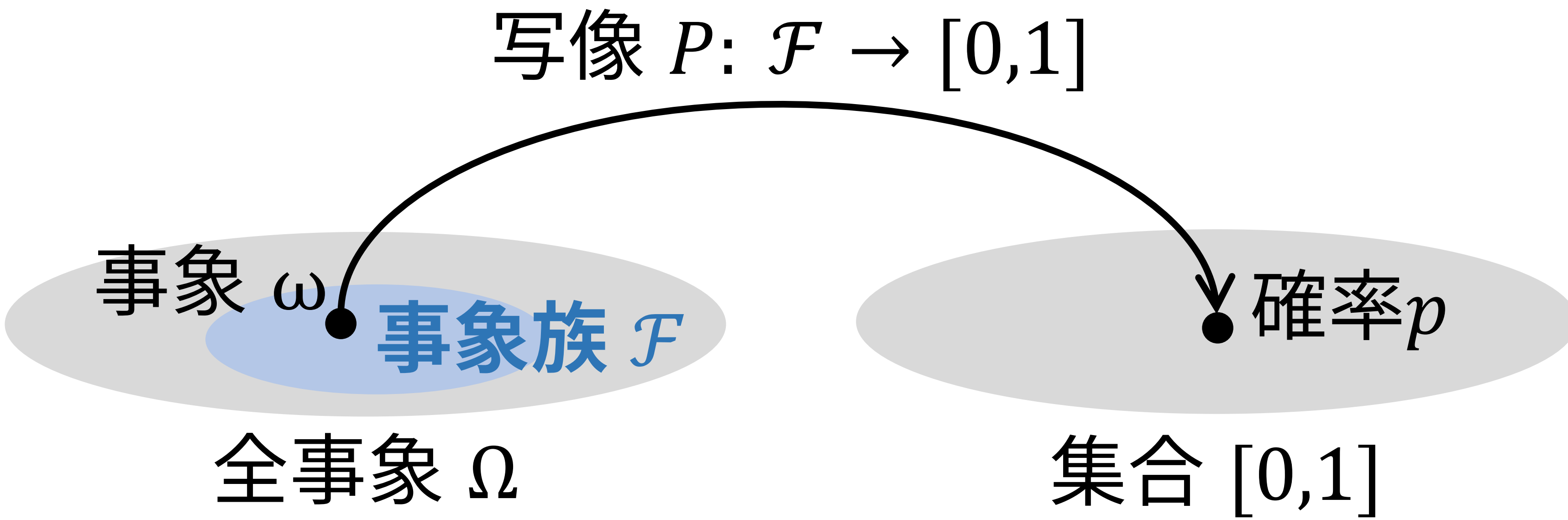
へのへのもへサイコロ
の
も
へ
全事象 Ω
面1: へ
面2: の
面3: へ
面4: の
面5: も
面6: へ
事象族 𝓕
へ
の
も
事象 ω
写像 P
集合[1,0]
1/2
1/3
1/6
確率 p

確率$\boldsymbol{p}$を定義するために
事象族$\mathcal{F}$が満たすべき<u>譲れない５条件</u>

確率pを定義するために
事象族$\mathcal{F}$が満たすべき<u>譲れない５条件</u>

1. 事象族$\mathcal{F}$は全事象Ωの部分集合である
$$\mathcal{F} \subseteq \Omega$$
2. 空集合$\emptyset$と全事象Ωは事象族$\mathcal{F}$の要素である
$$\emptyset, \Omega \in \mathcal{F}$$
3. 事象族$\mathcal{F}$は空集合$\emptyset$であってはならない
$$\mathcal{F} \neq \emptyset$$
4. 事象ωを要素に持つ時、その余事象$\overline{\omega}$も要素である
$$\omega \in \mathcal{F} \quad \Rightarrow \quad \overline{\omega} \in \mathcal{F}$$
5. 要素同士の和集合も要素である
$$\mathcal{F} = \{\omega_1, \ldots, \omega_n\} \;\Rightarrow\; \bigcup_{i \geq 1} \omega_i \in \mathcal{F}$$



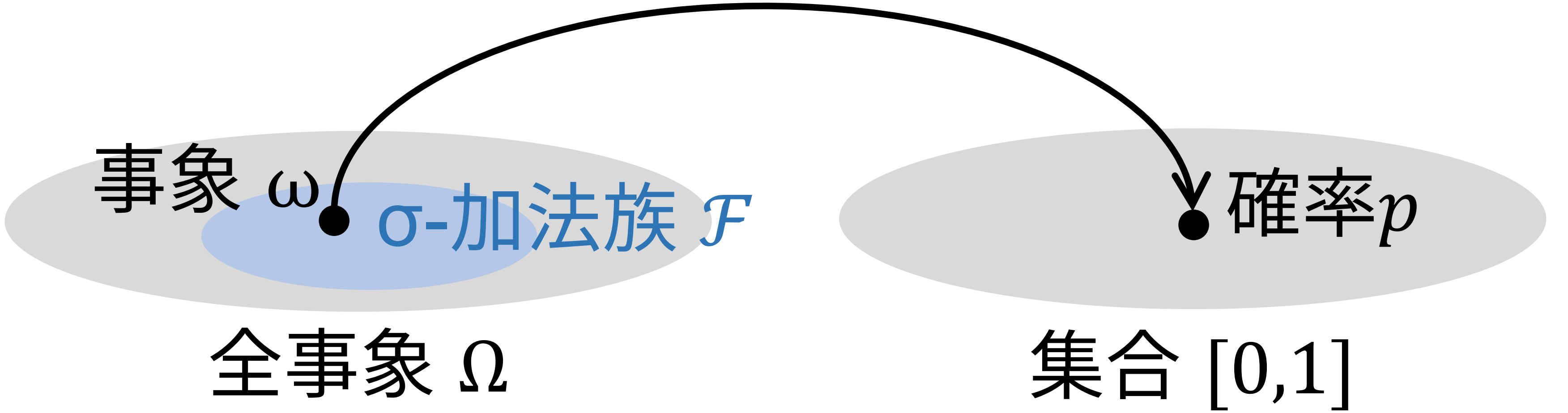


事象族$\mathcal{F}$が<u>譲れない５条件</u>を満たす

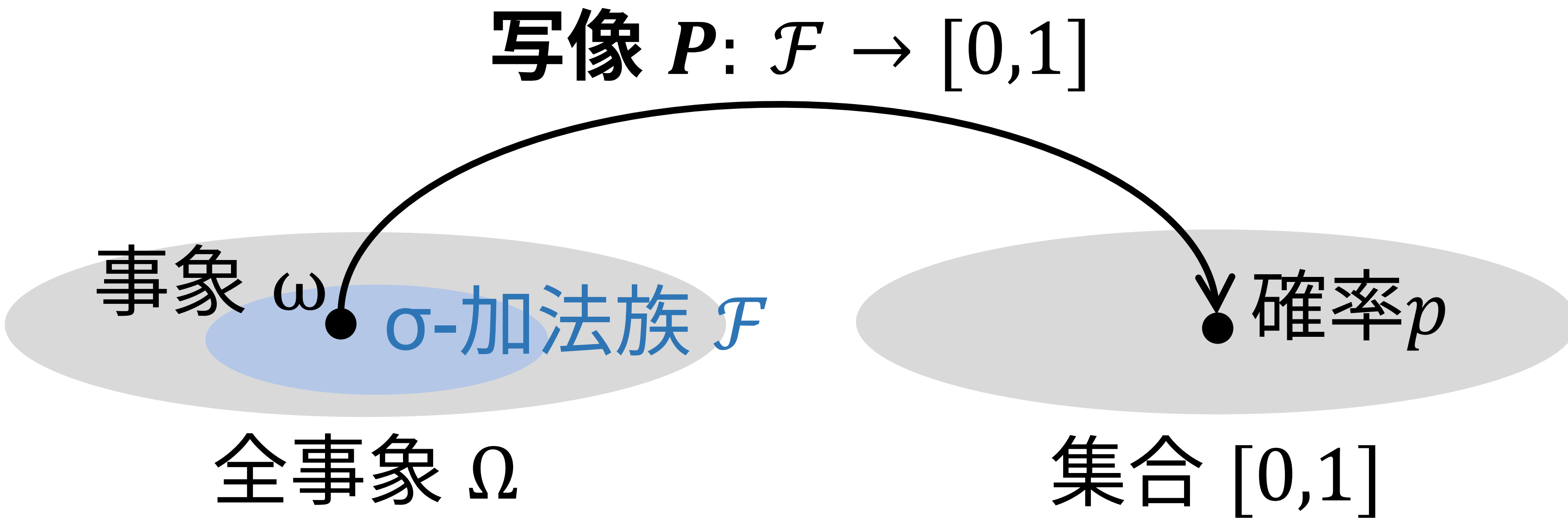
→ **σ-加法族**$\mathcal{F}$と呼ぶ（σ-加法性を満たす）
σ-additivity

確率pが定義される**事象族**$\mathcal{F}$は**σ-加法性**を満たさなければならない。

写像 $\boldsymbol{P}\colon \mathcal{F} \to [0,1]$

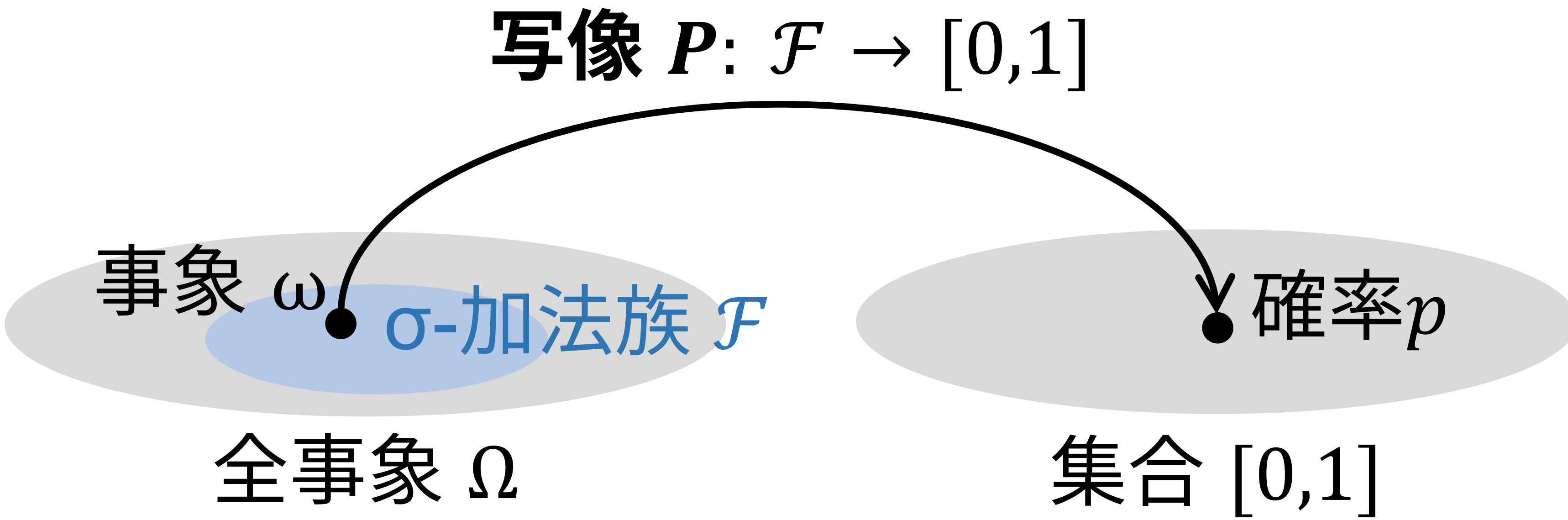


写像 $\boldsymbol{P}$が満たすべき<u>譲れない３条件</u>



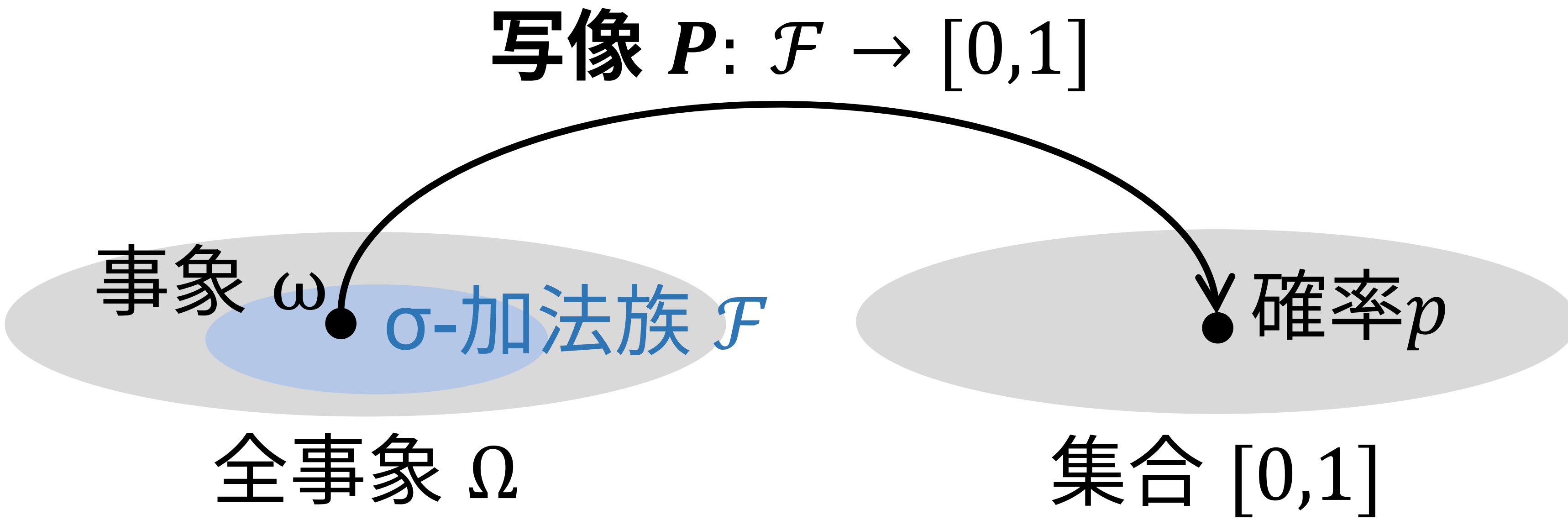
写像 $\boldsymbol{P}$が満たすべき譲れない３条件

1. 写像P: $\mathcal{F} \to [0,1]$もしくはP: $\omega \subseteq \Omega \mapsto p \in [0,1]$



写像 $\boldsymbol{P}$が満たすべき<u>譲れない３条件</u>

1. 写像P: $\mathcal{F} \to [0,1]$もしくはP: $\omega \subseteq \Omega \mapsto p \in [0,1]$
2. 全事象の確率 $P(\Omega) = 1$



写像 Pが満たすべき<u>譲れない３条件</u>

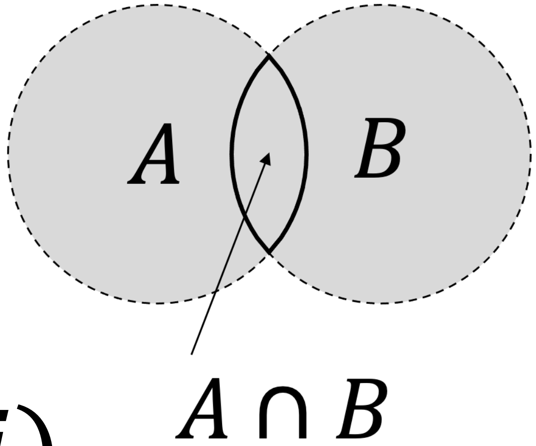
1. 写像P: $\mathcal{F} \to [0,1]$もしくはP: $\omega \subseteq \Omega \mapsto p \in [0,1]$
2. 全事象の確率 $P(\Omega) = 1$
3. 可算加法性を満たす

可算加法性を満たす

$$\forall\left(\omega_i \cap \omega_j\right) = \emptyset \quad (i \neq j)$$

$$\Rightarrow P\left(\bigcup_{i\geq 1} \omega_i\right) = \sum_{i\geq 1} P(\omega_i)$$

可算加法性を満たす

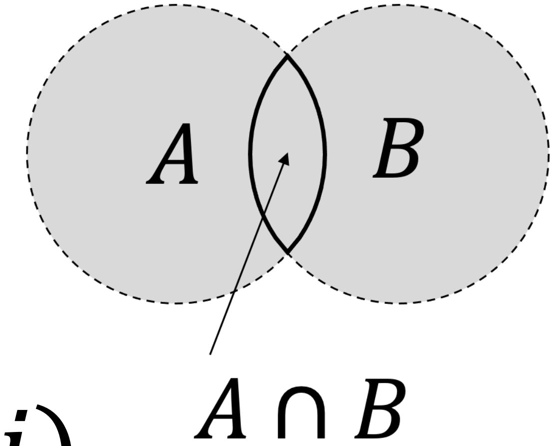


互いに排反ならば、

$$\forall\left(\omega_i \cap \omega_j\right) = \emptyset \quad (i \neq j)$$

$$\Rightarrow P\left(\bigcup_{i \geq 1} \omega_i\right) = \sum_{i \geq 1} P(\omega_i)$$

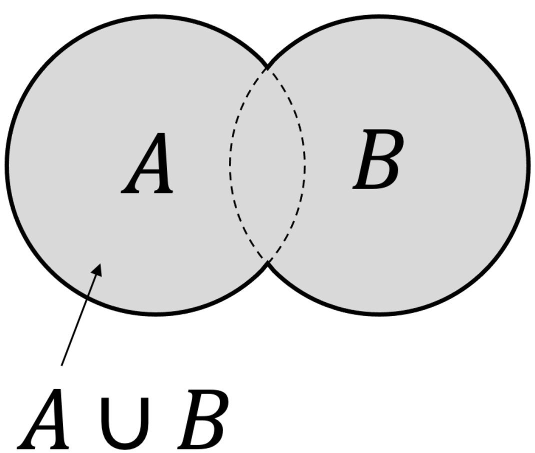
可算加法性を満たす



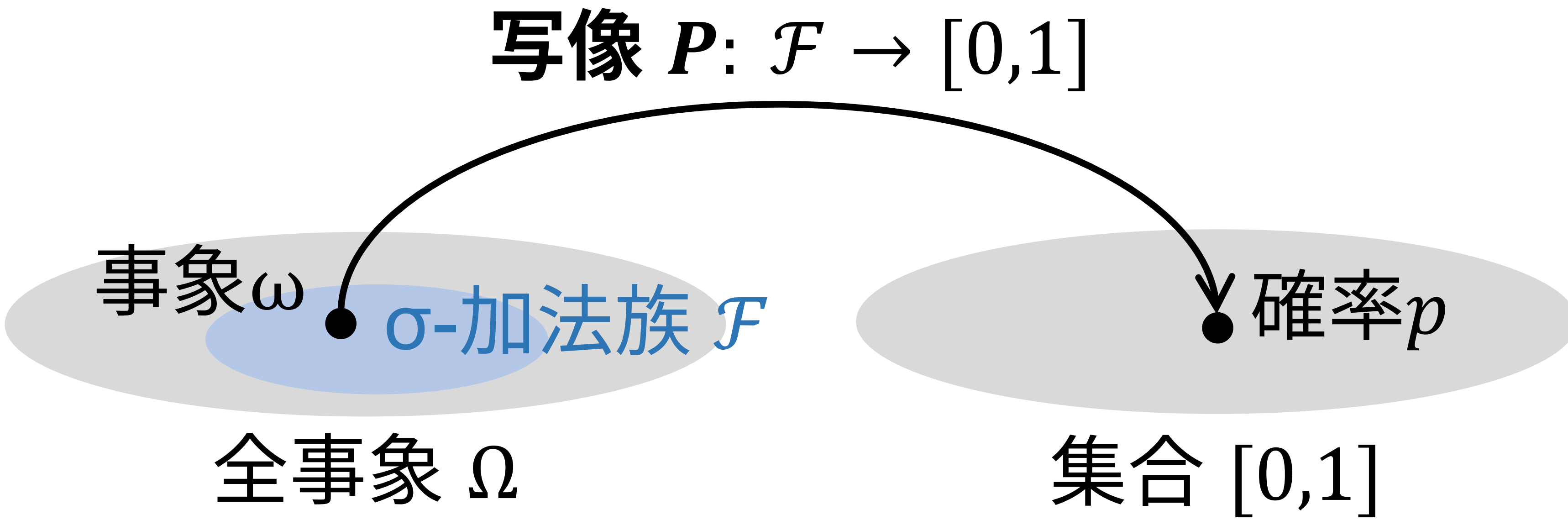
互いに排反ならば、

$$\forall\left(\omega_i \cap \omega_j\right) = \emptyset \quad (i \neq j)$$

$$\Rightarrow P\left(\bigcup_{i \geq 1} \omega_i\right) = \sum_{i \geq 1} P(\omega_i)$$

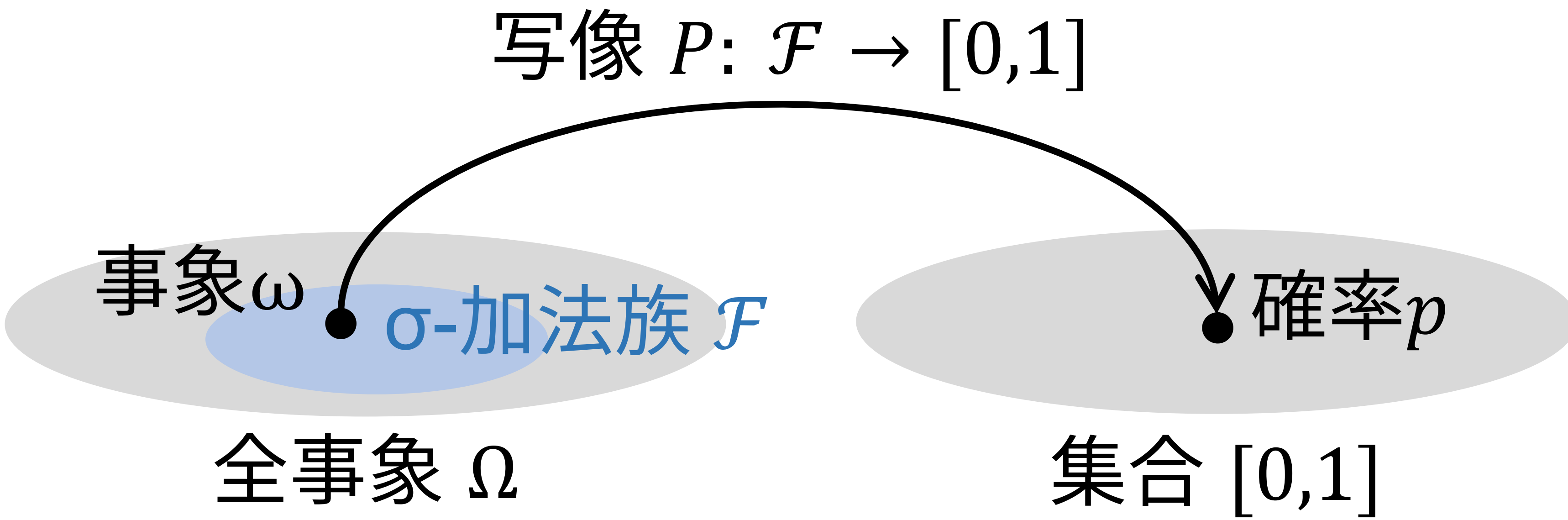


和事象の確率は、　事象の確率の和



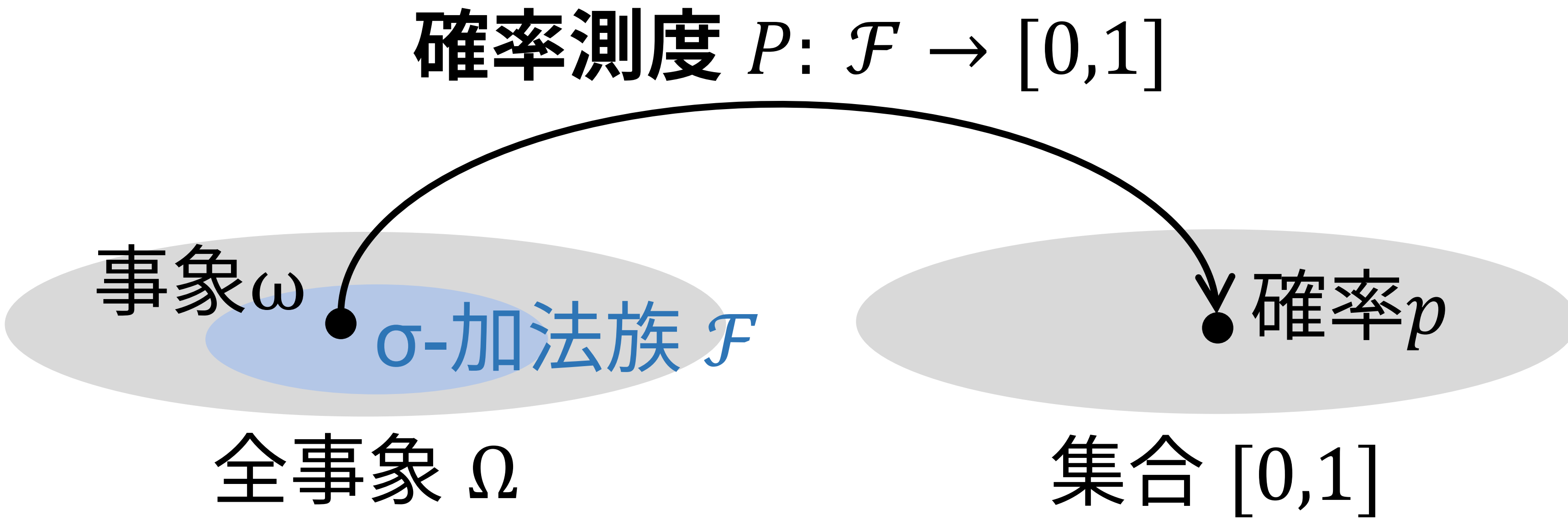
写像 $\boldsymbol{P}$が満たすべき<u>譲れない３条件</u>

1. 写像P: $\mathcal{F} \to [0,1]$もしくはP: $\omega \subseteq \Omega \mapsto p \in [0,1]$
2. 全事象の確率$P(\Omega) = 1$
3. 可算加法性を満たす



事象族$\mathcal{F}$が満たすべき$\underbrace{\text{譲れない５条件}}_{\text{σ−加法性}}$

写像 Pが満たすべき$\underbrace{\text{譲れない３条件}}_{\text{コルモゴロフの公理系}}$



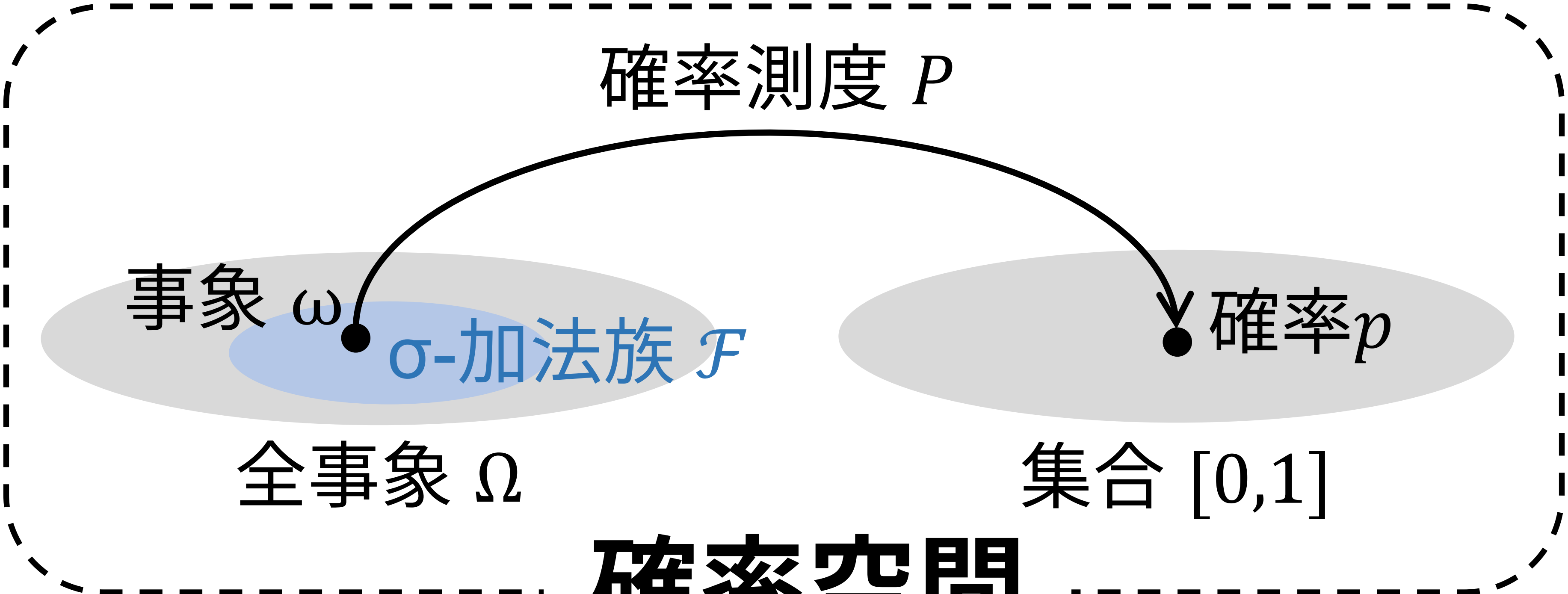
事象族$\mathcal{F}$が<u>譲れない５条件</u>を満たす

→ **σ-加法族**$\mathcal{F}$と呼ぶ（σ-加法性を満たす）

写像Pが<u>譲れない３条件</u>を満たす

→ **確率測度**Pと呼ぶ

probability measure



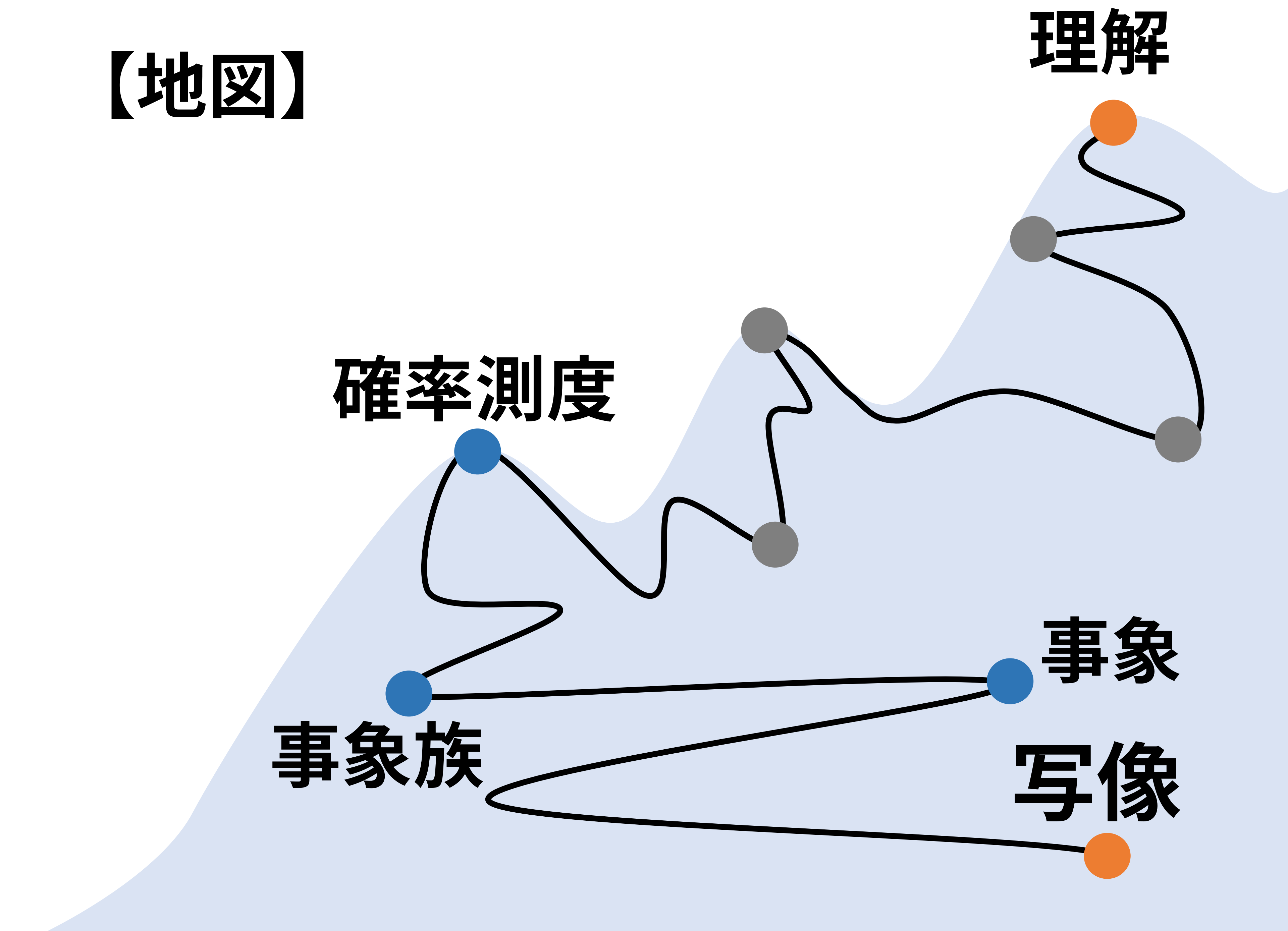
確率空間

$$\mathcal{P}[\Omega, \mathcal{F}, P]$$

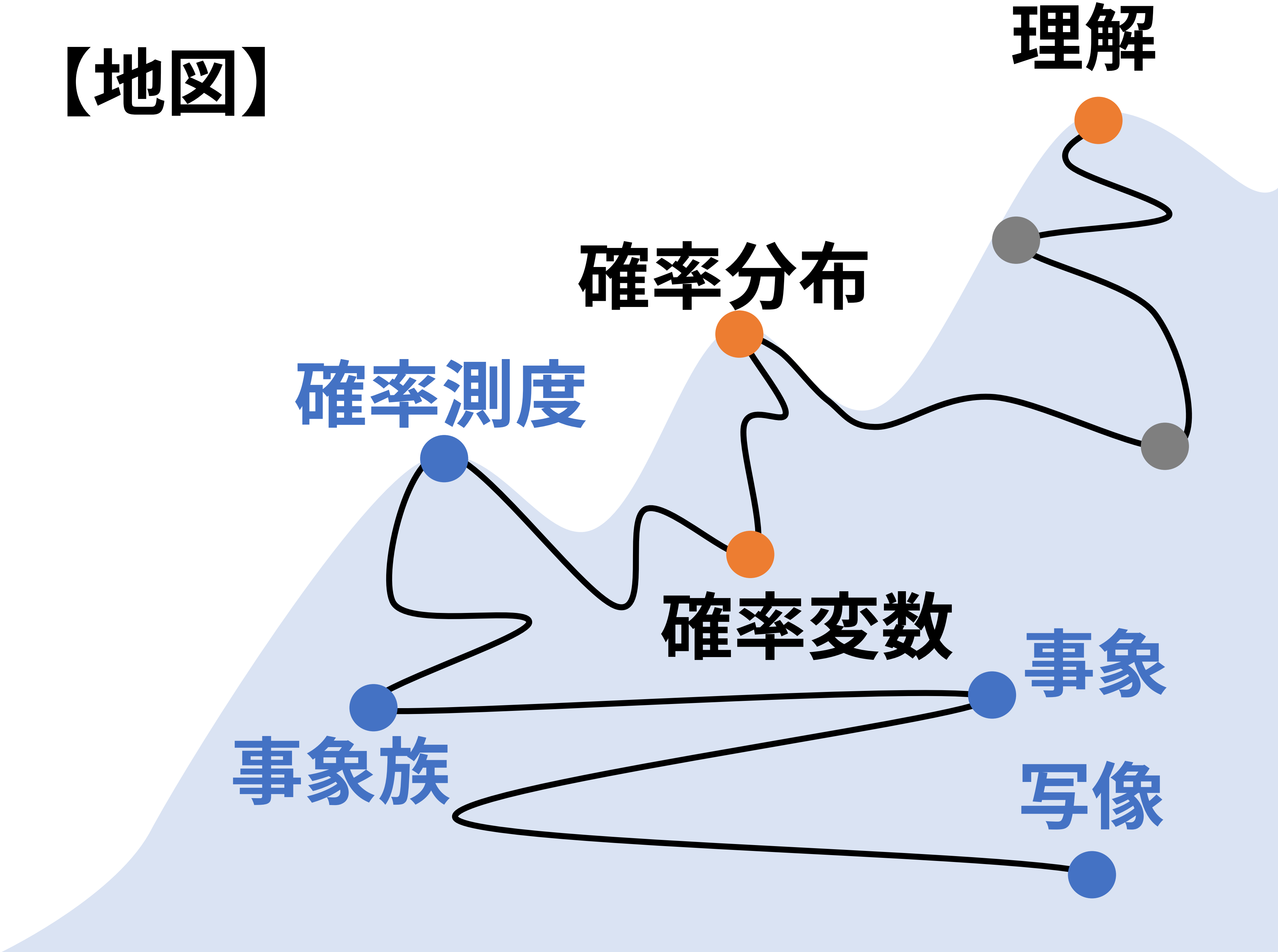
全事象Ω: 起こりうる全ての事象を網羅した集合

事象族$\mathcal{F}$: σ-加法性を満たし写像のchannelを規定する

確率測度P: 事象ωに対し確率pを対応づける写像

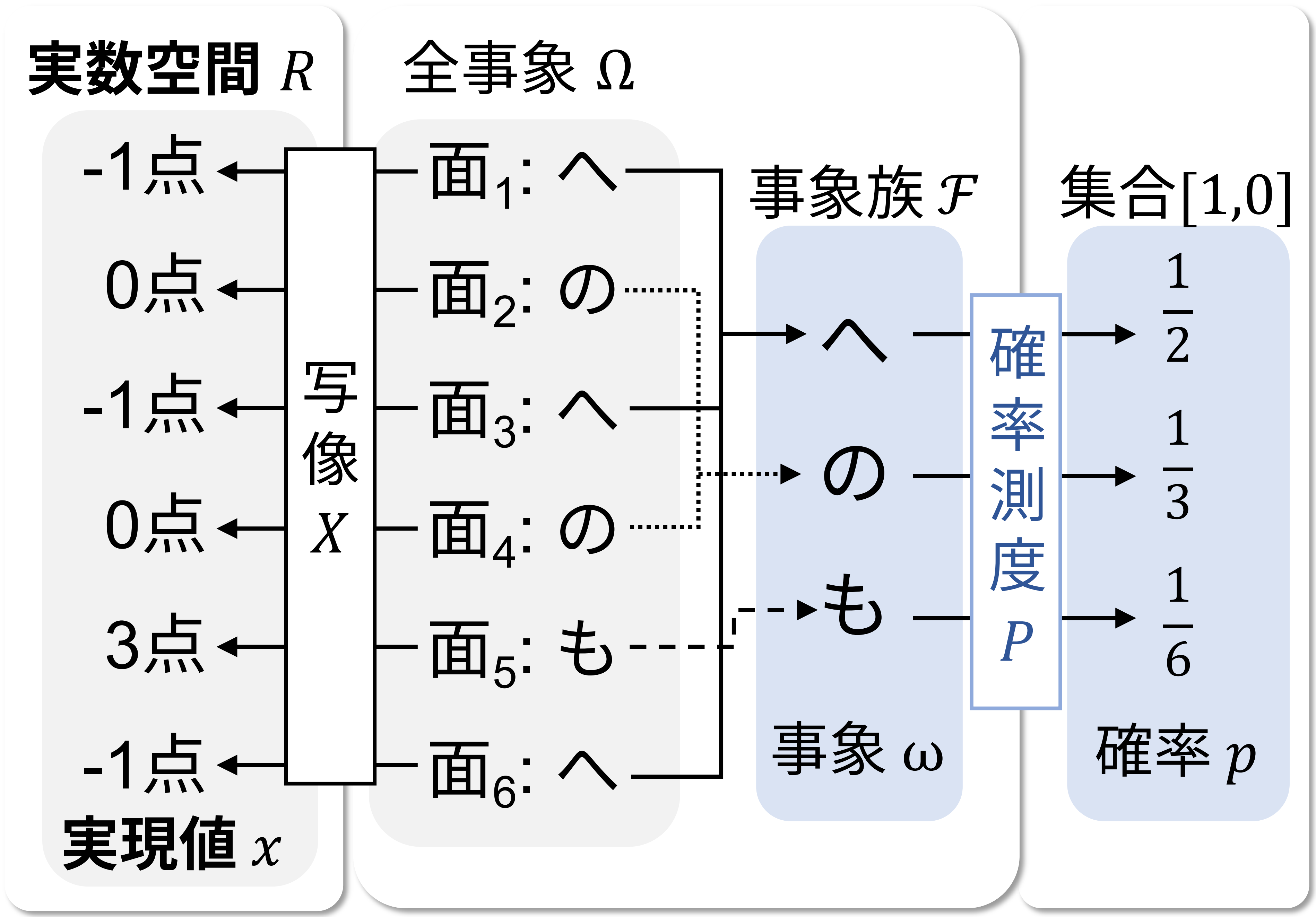

【地図】
理解
確率測度
事象
事象族
写像

【地図】

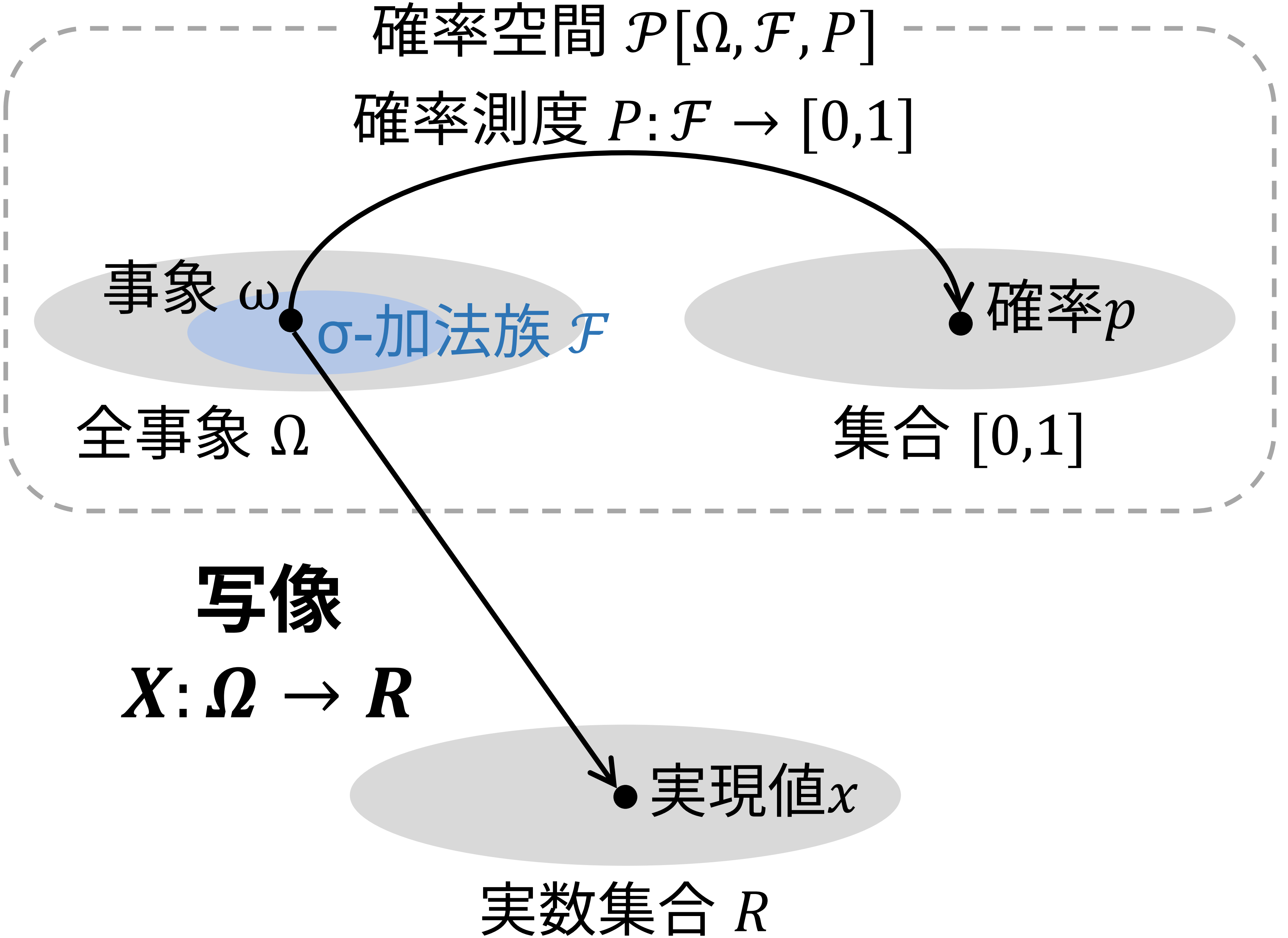


かくりつへんすう　しゃぞう

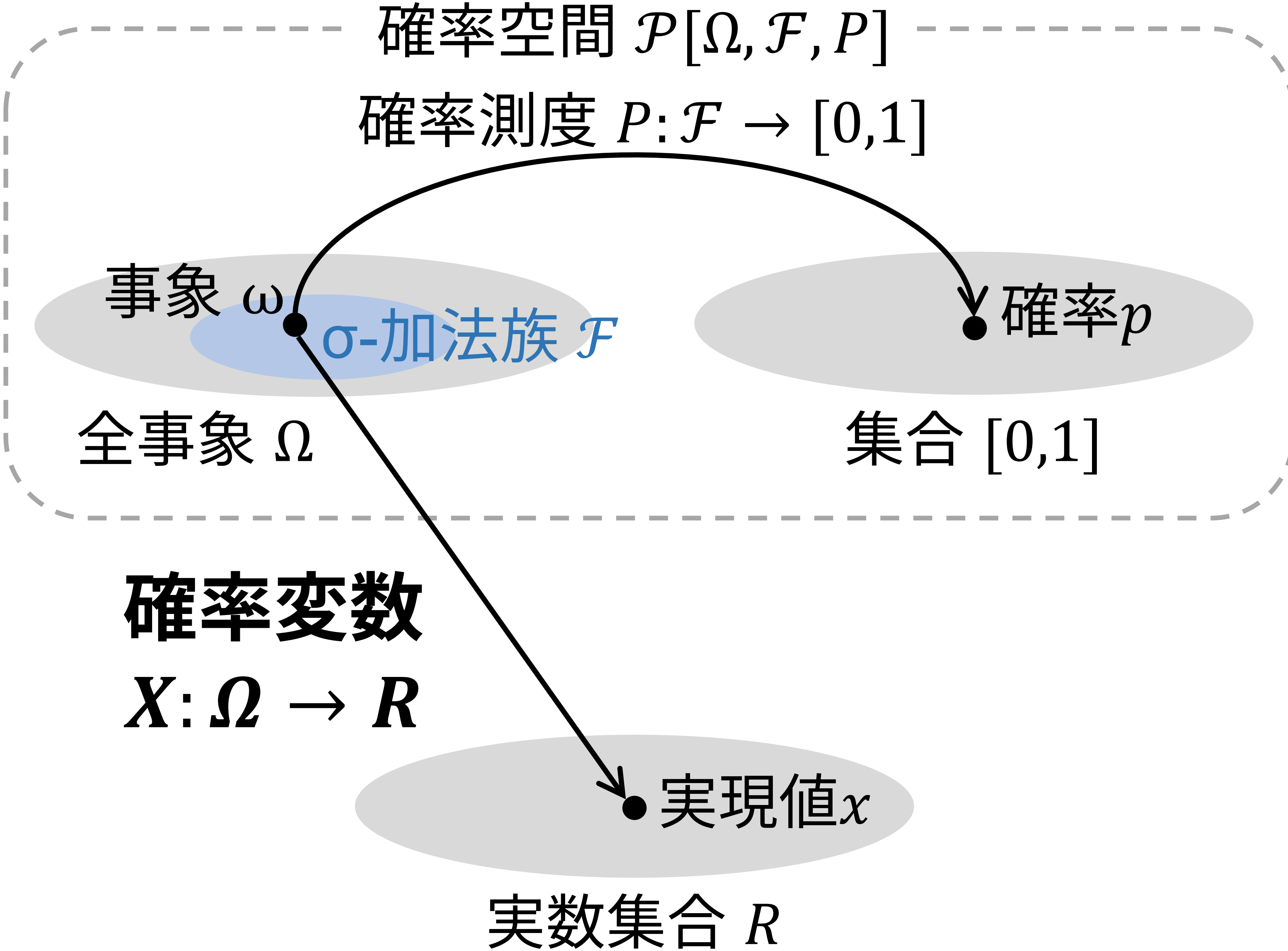
確率変数は写像



実数空間 R
全事象 Ω
-1点
0点
-1点
0点
3点
-1点
実現値 x
写像 X
面$_1$: へ
面$_2$: の
面$_3$: へ
面$_4$: の
面$_5$: も
面$_6$: へ
事象族 $\mathcal{F}$
へ
の
も
事象 ω
確率測度 P
集合[1,0]
$\frac{1}{2}$
$\frac{1}{3}$
$\frac{1}{6}$
確率 p

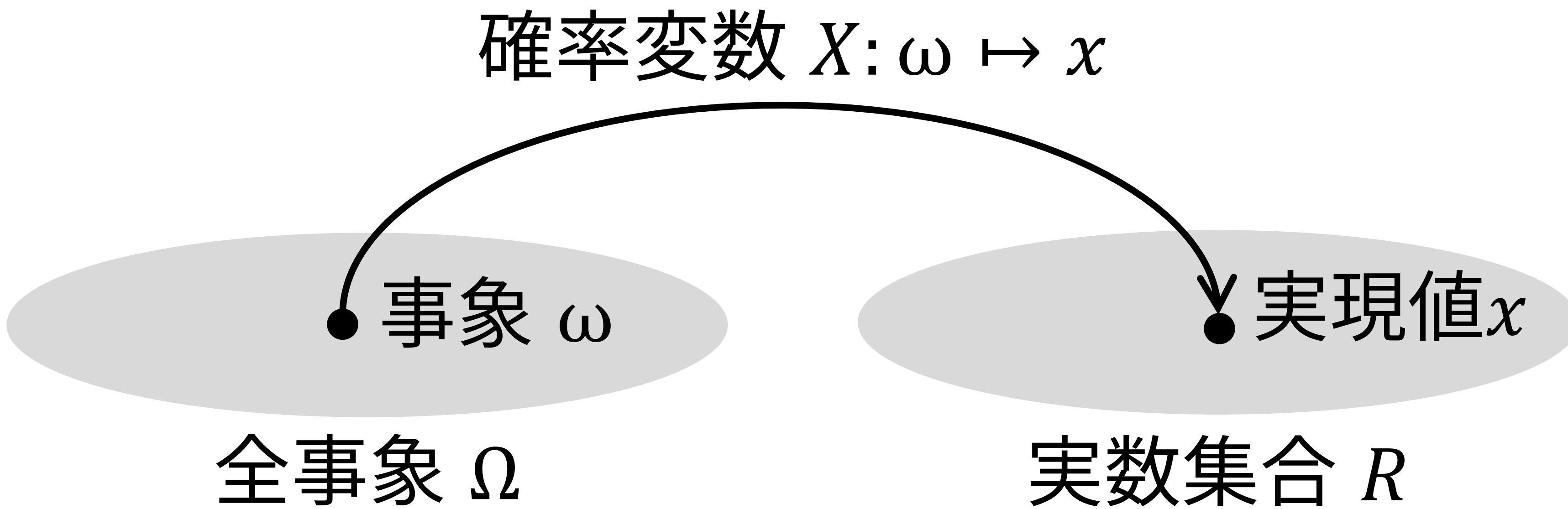


確率空間 $\mathcal{P}[\Omega, \mathcal{F}, P]$
確率測度 $P: \mathcal{F} \to [0,1]$
事象 ω
σ-加法族 $\mathcal{F}$
全事象 Ω
確率p
集合 [0,1]
写像
$\boldsymbol{X}: \boldsymbol{\Omega} \to \boldsymbol{R}$
実現値x
実数集合 R

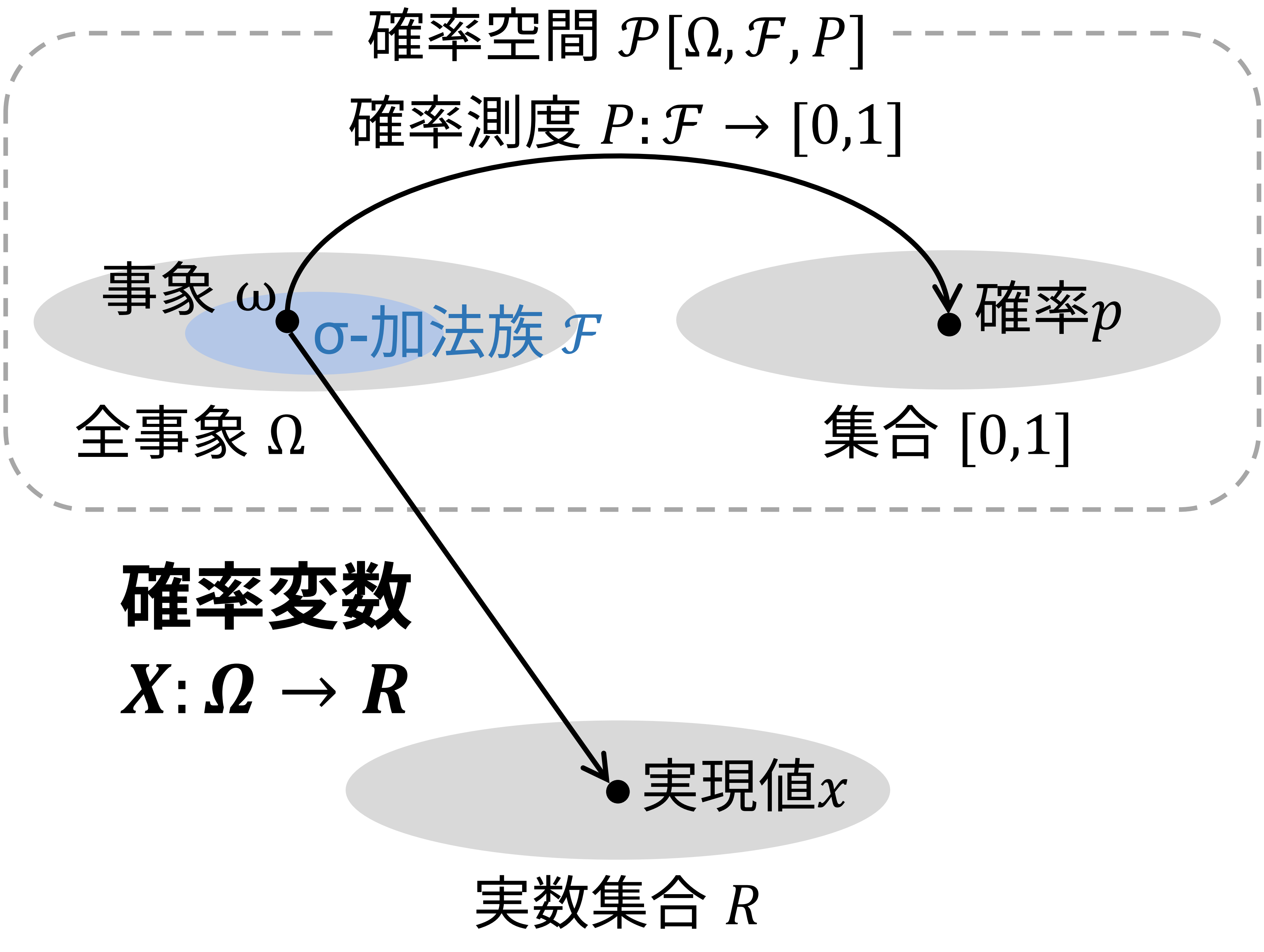


確率空間 $\mathcal{P}[\Omega, \mathcal{F}, P]$
確率測度 $P: \mathcal{F} \to [0,1]$
事象 ω
σ-加法族 $\mathcal{F}$
確率p
全事象 Ω
集合 [0,1]
確率変数
$\boldsymbol{X: \Omega \to R}$
実現値x
実数集合 R

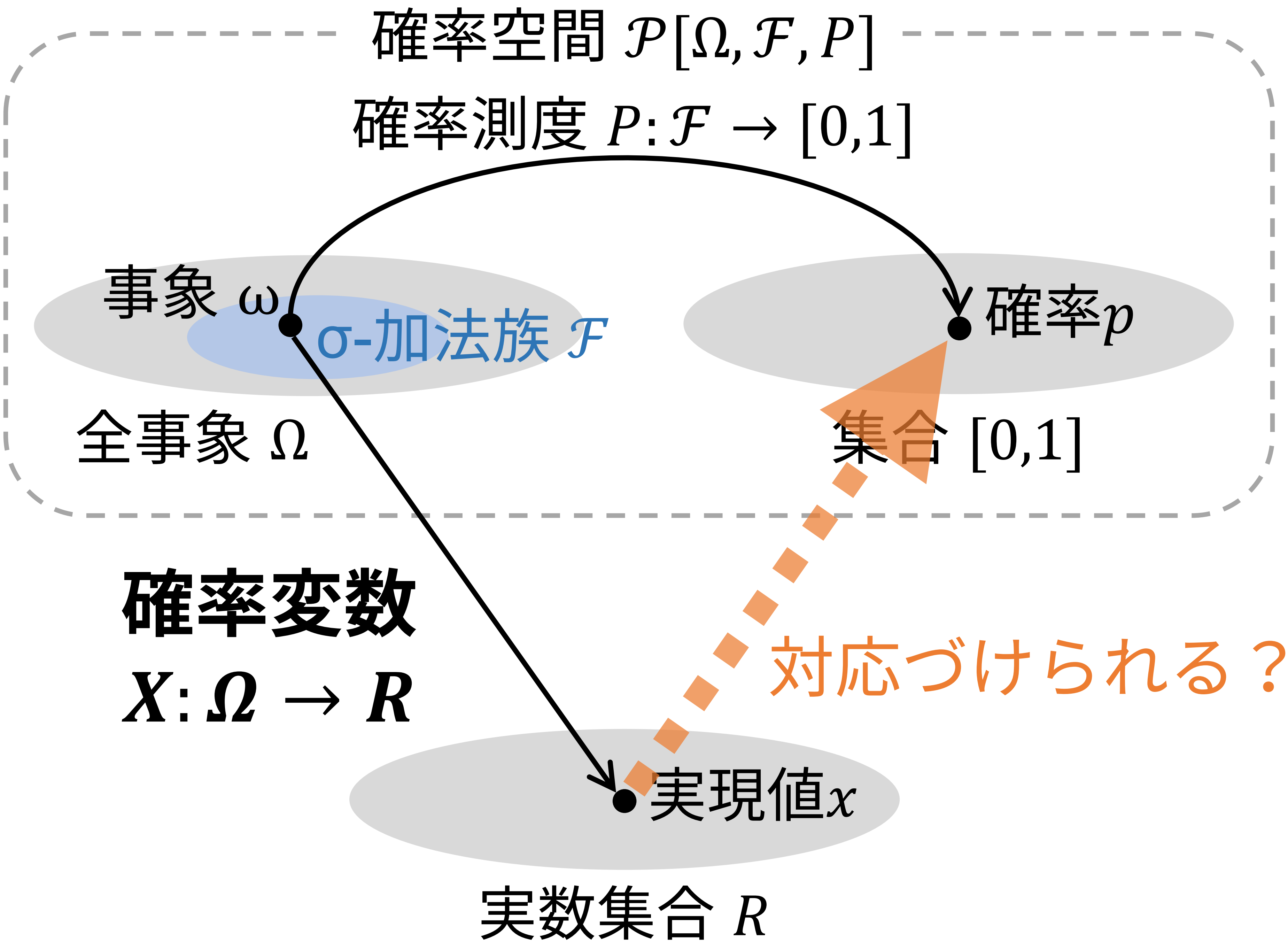
確率変数 は 写像



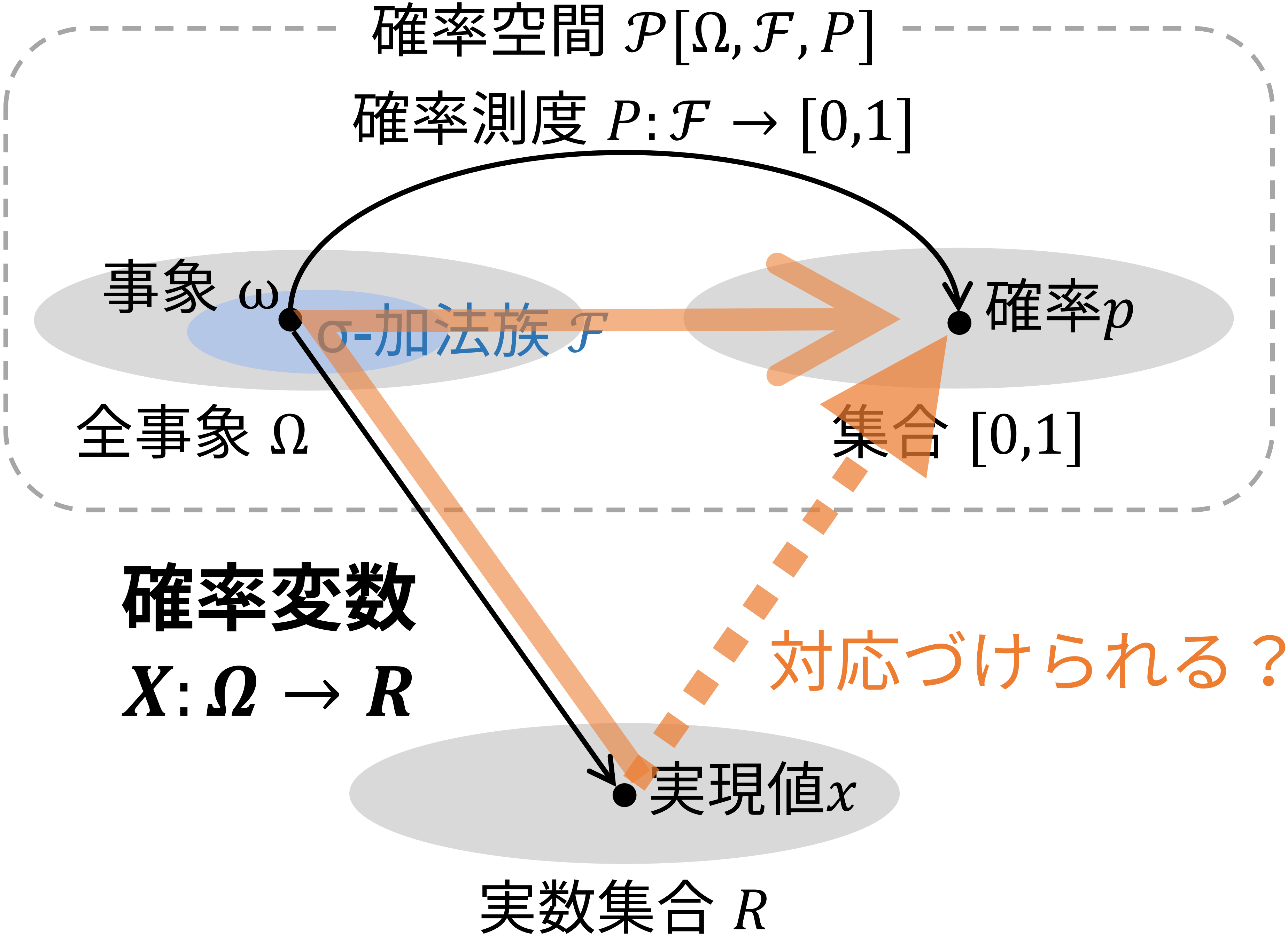
全事象Ωの要素(事象ω)を実数集合Rのただ1つの要素(実現値x)に対応づける写像



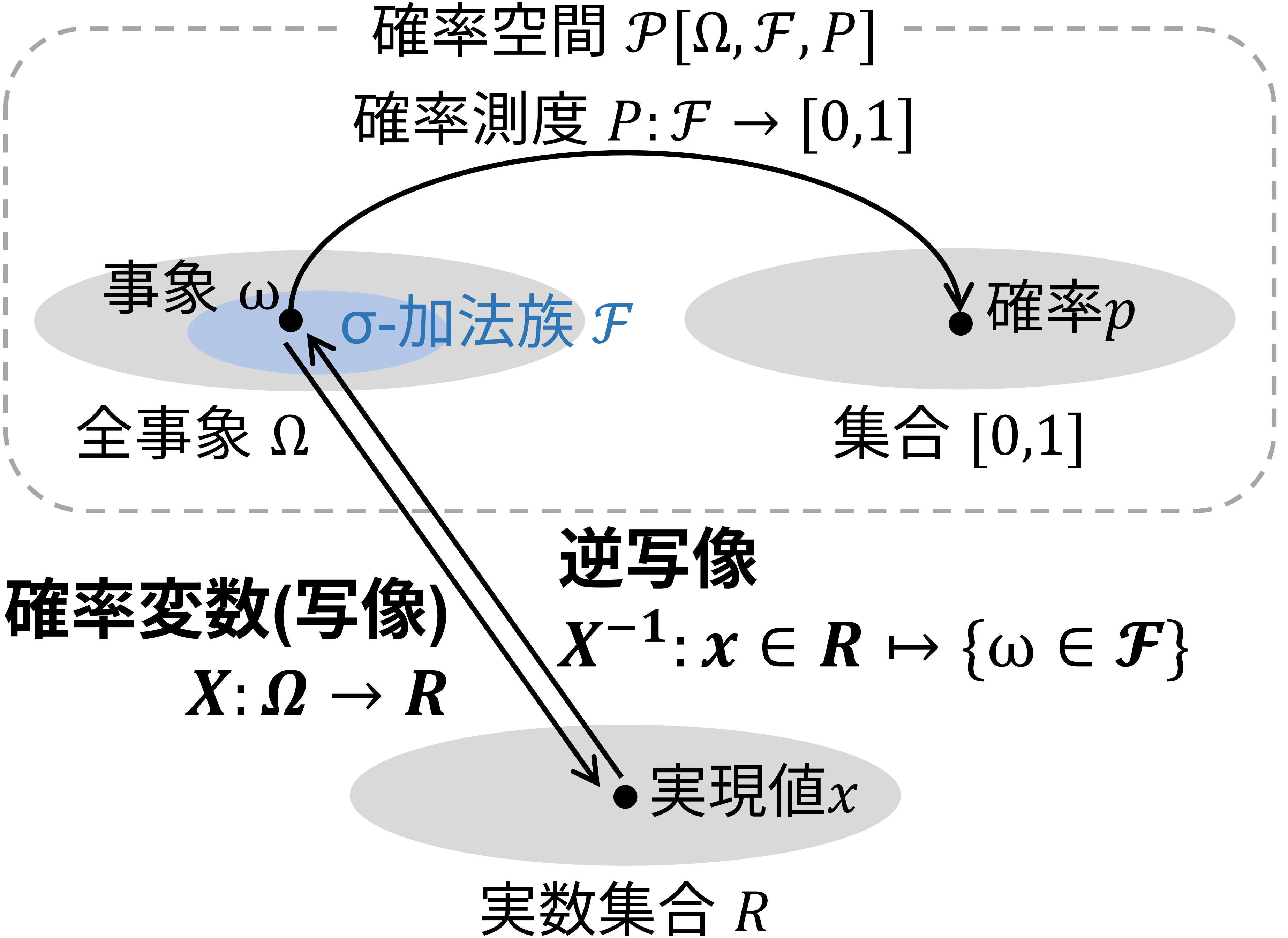
確率空間 $\mathcal{P}[\Omega, \mathcal{F}, P]$
確率測度 $P: \mathcal{F} \to [0,1]$
事象 ω
σ-加法族 $\mathcal{F}$
確率p
全事象 Ω
集合 [0,1]
確率変数
$\boldsymbol{X}: \boldsymbol{\Omega} \to \boldsymbol{R}$
実現値x
実数集合 R

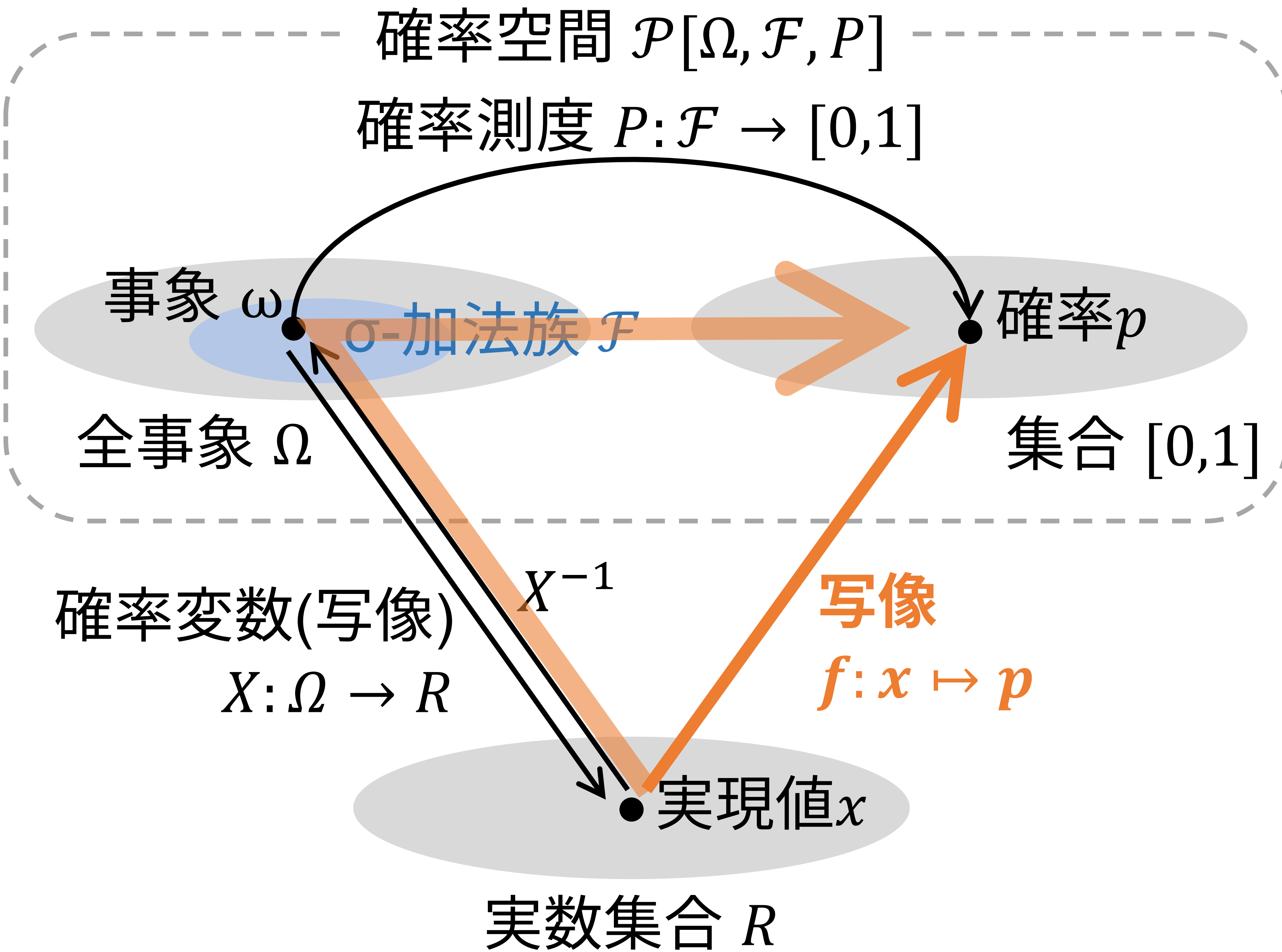


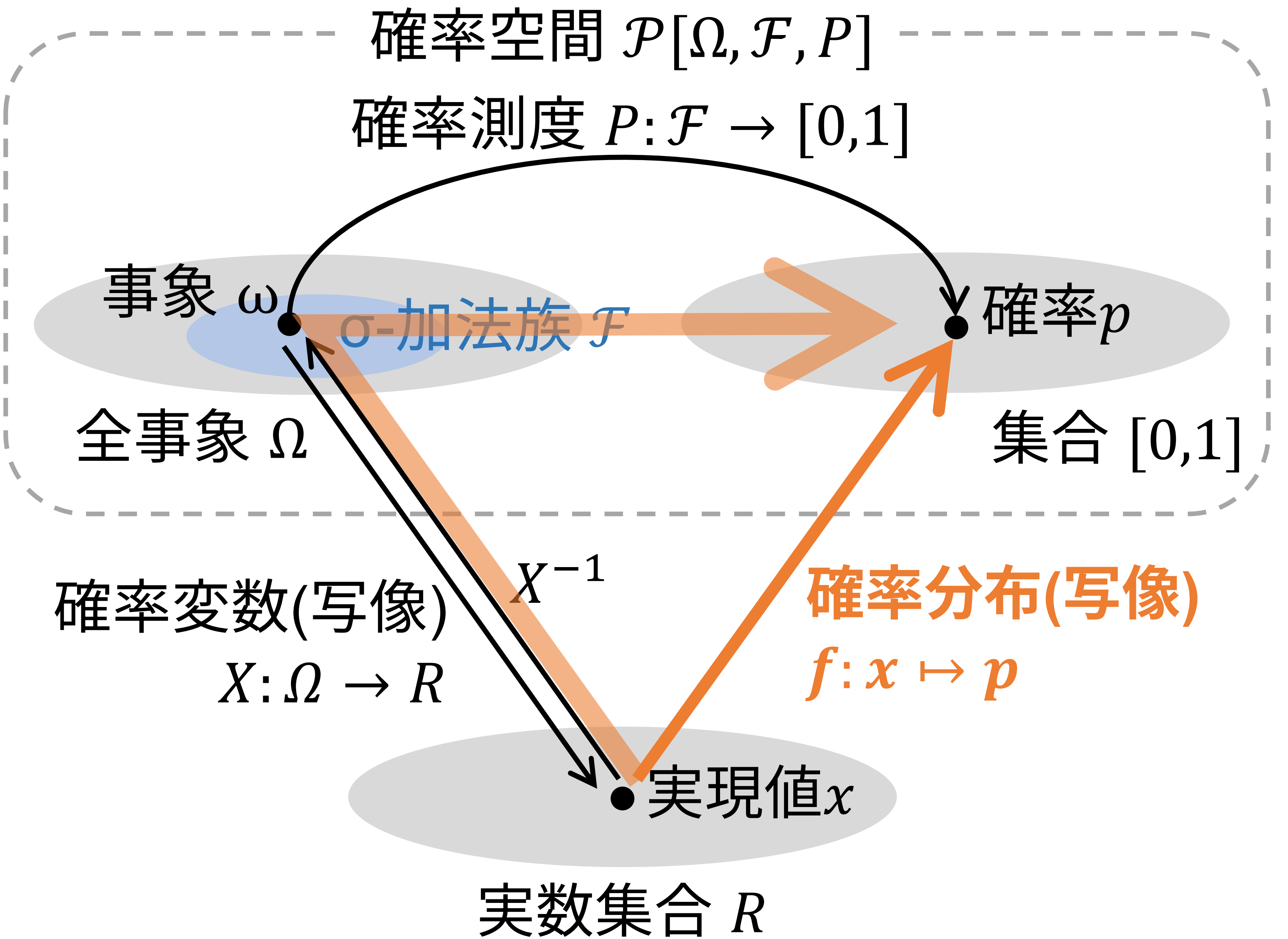
確率空間 $\mathcal{P}[\Omega, \mathcal{F}, P]$
確率測度 $P: \mathcal{F} \to [0,1]$
事象 ω
σ-加法族 $\mathcal{F}$
確率p
全事象 Ω
集合 [0,1]
確率変数
$X: \Omega \to R$
対応づけられる？
実現値x
実数集合 R



確率空間 $\mathcal{P}[\Omega, \mathcal{F}, P]$
確率測度 $P: \mathcal{F} \to [0,1]$
事象 ω
σ-加法族 $\mathcal{F}$
確率p
全事象 Ω
集合 [0,1]
確率変数
$X: \Omega \to R$
対応づけられる？
実現値x
実数集合 R


確率空間 $\mathcal{P}[\Omega, \mathcal{F}, P]$
確率測度 $P: \mathcal{F} \to [0,1]$
事象 ω
σ-加法族 $\mathcal{F}$
確率p
全事象 Ω
集合 [0,1]
逆写像
$X^{-1}: x \in R \mapsto \{\omega \in \mathcal{F}\}$
確率変数(写像)
$X: \Omega \to R$
実現値x
実数集合 R


確率空間 $\mathcal{P}[\Omega, \mathcal{F}, P]$
確率測度 $P: \mathcal{F} \to [0,1]$
事象 ω
σ-加法族 $\mathcal{F}$
確率p
全事象 Ω
集合 [0,1]
X^{-1}
確率変数(写像)
$X: \Omega \to R$
写像
$f: x \mapsto p$
実現値x
実数集合 R

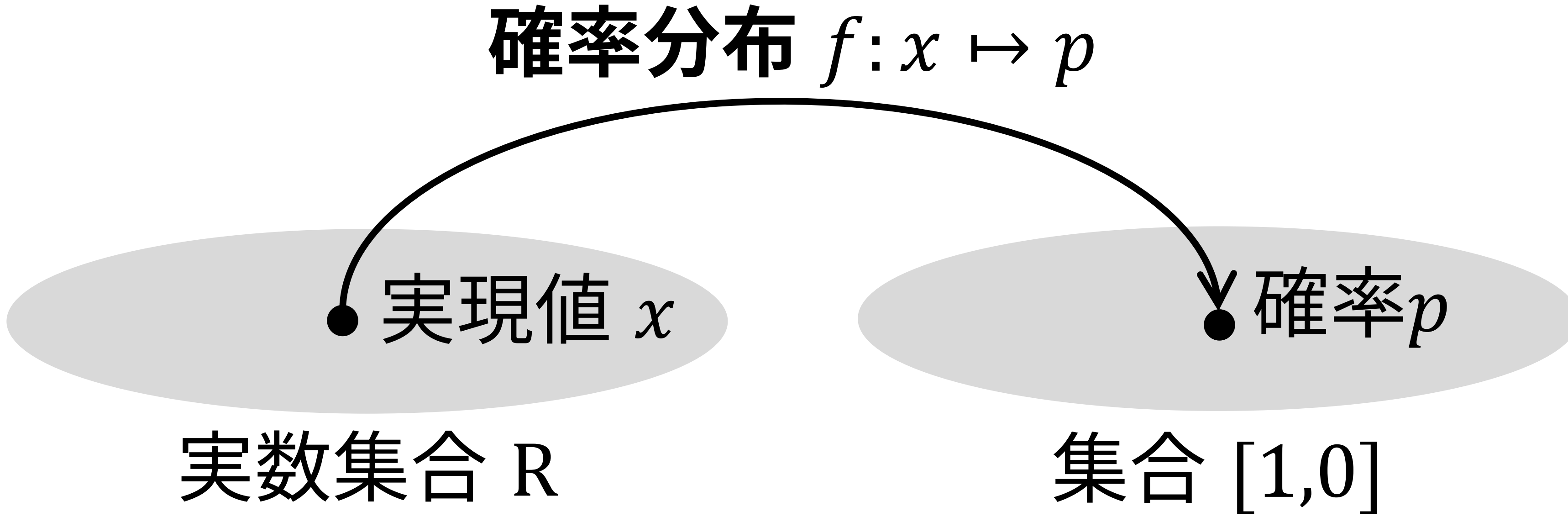


確率空間 $\mathcal{P}[\Omega, \mathcal{F}, P]$
確率測度 $P: \mathcal{F} \to [0,1]$
事象 ω
σ-加法族 $\mathcal{F}$
確率p
全事象 Ω
集合 [0,1]
X^{-1}
確率変数(写像)
$X: \Omega \to R$
確率分布(写像)
$f: x \mapsto p$
実現値x
実数集合 R

かくりつぶんぷ　しゃぞう

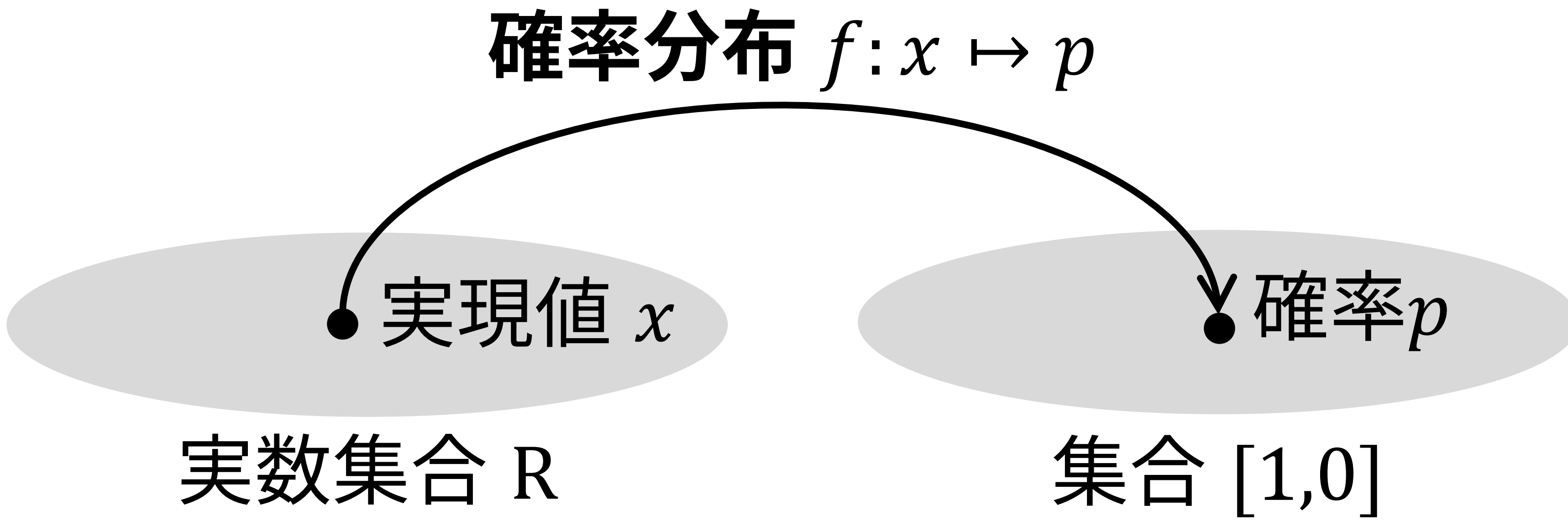
確率分布は写像

確率分布 は 写像



実数集合Rの要素(実現値x)を集合[0,1]のただ1つの要素(確率p)に対応づける規則

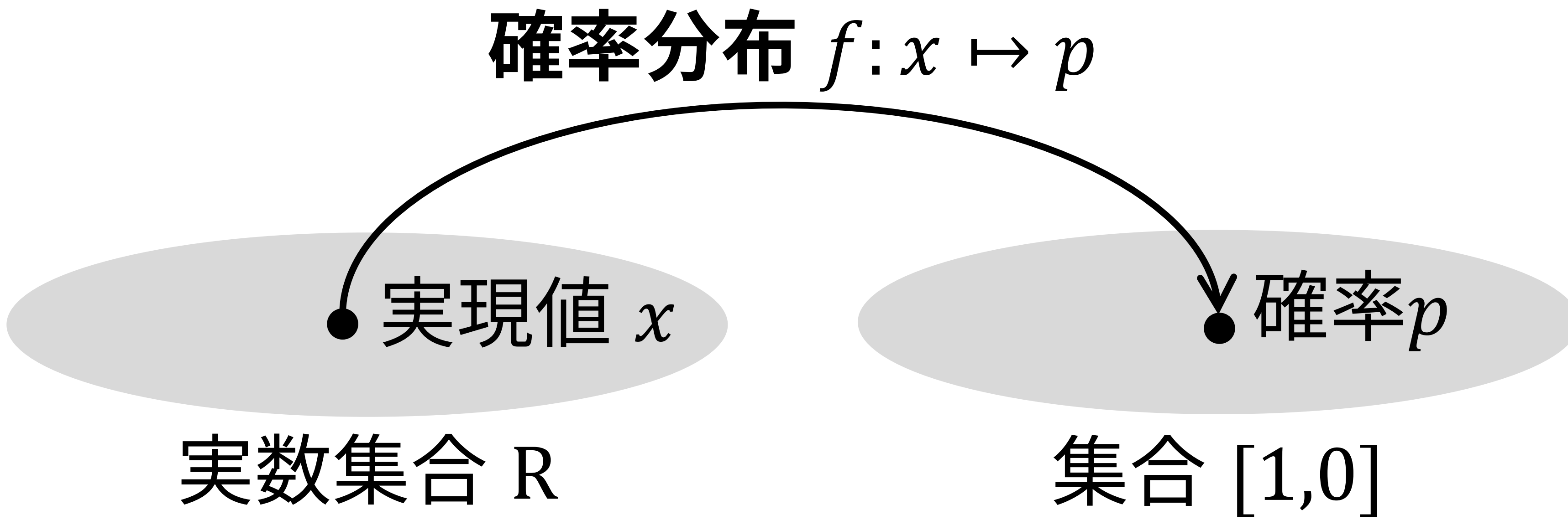
確率分布 は 写像



実数集合Rの要素(実現値x)を集合[0,1]のただ1つの要素(確率p)に対応づける規則

ただし、

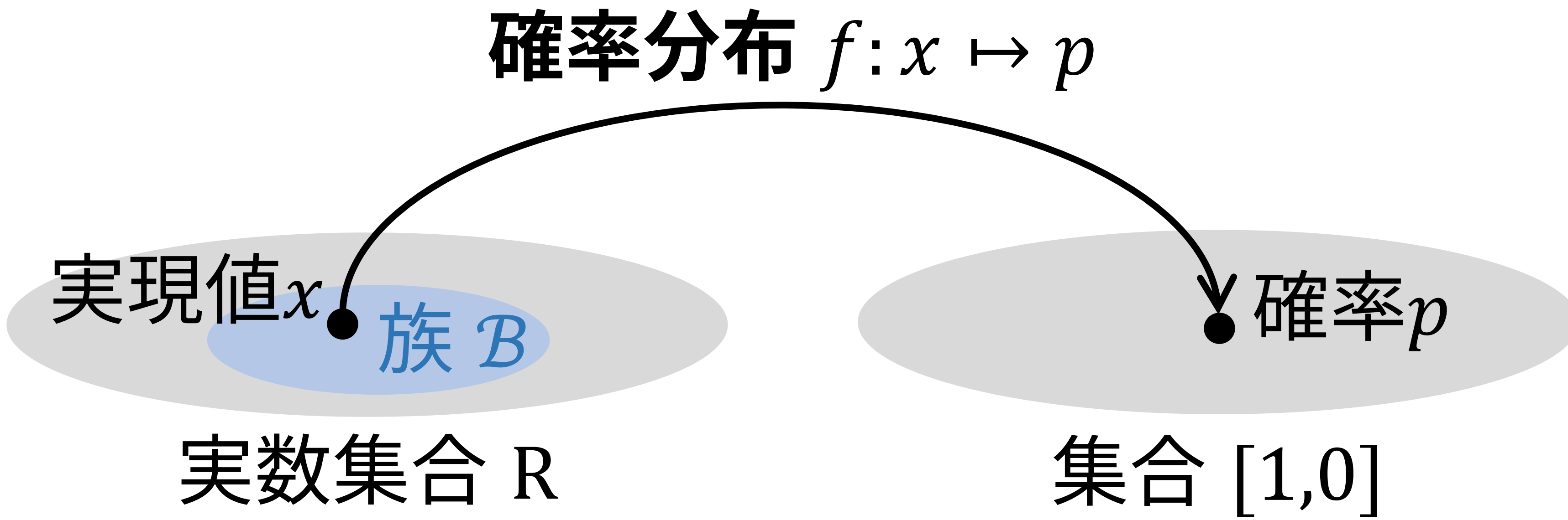
確率分布 は 写像



実数集合Rの要素(実現値x)を集合[0,1]のただ1つの要素(確率p)に対応づける規則

ただし、channelの規定が必要

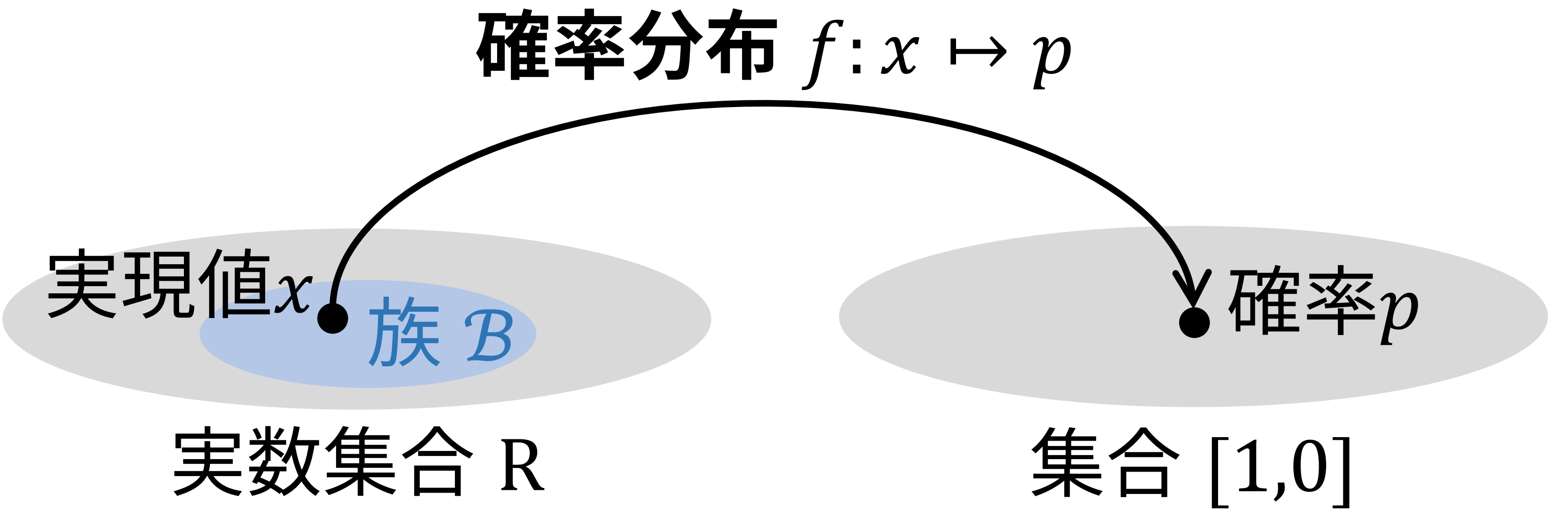
確率分布 は 写像



実数集合Rの要素(実現値x)を集合$[0,1]$のただ1つの要素(確率p)に対応づける規則

ただし、channel(族$\mathcal{B}$)の規定が必要

確率分布 は 写像



族$\mathcal{B}$が満たすべき<u>譲れない５条件</u>

確率pを定義するために
事象族$\mathcal{B}$が満たすべき<u>譲れない５条件</u>

1. 事象族$\mathcal{F}$は全事象Ωの部分集合である

$$\mathcal{F} \subseteq \Omega$$

2. 空集合$\emptyset$と全事象Ωは事象族$\mathcal{F}$の要素である

$$\emptyset, \Omega \in \mathcal{F}$$

3. 事象族$\mathcal{F}$は空集合$\emptyset$であってはならない

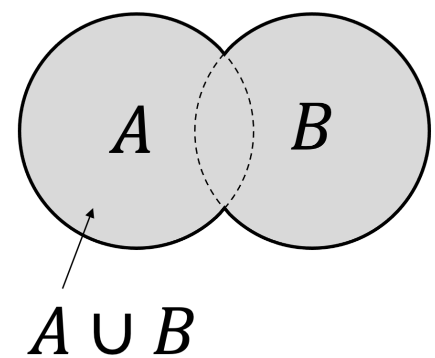
$$\mathcal{F} \neq \emptyset$$

4. 事象ωを要素に持つ時、その余事象$\overline{\omega}$も要素である

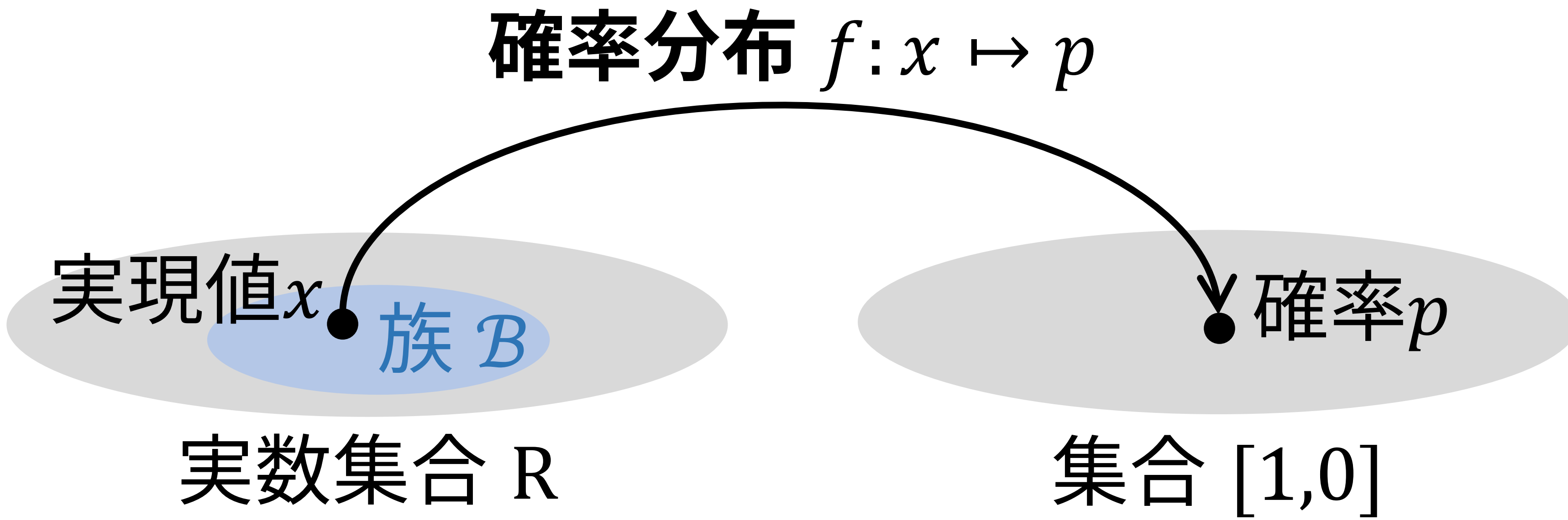
$$\omega \in \mathcal{F} \quad \Rightarrow \quad \overline{\omega} \in \mathcal{F}$$

5. 要素同士の和集合も要素である

$$\mathcal{F} = \{\omega_1, \ldots, \omega_n\} \ \Rightarrow \ \bigcup_{i \geq 1} \omega_i \in \mathcal{F}$$



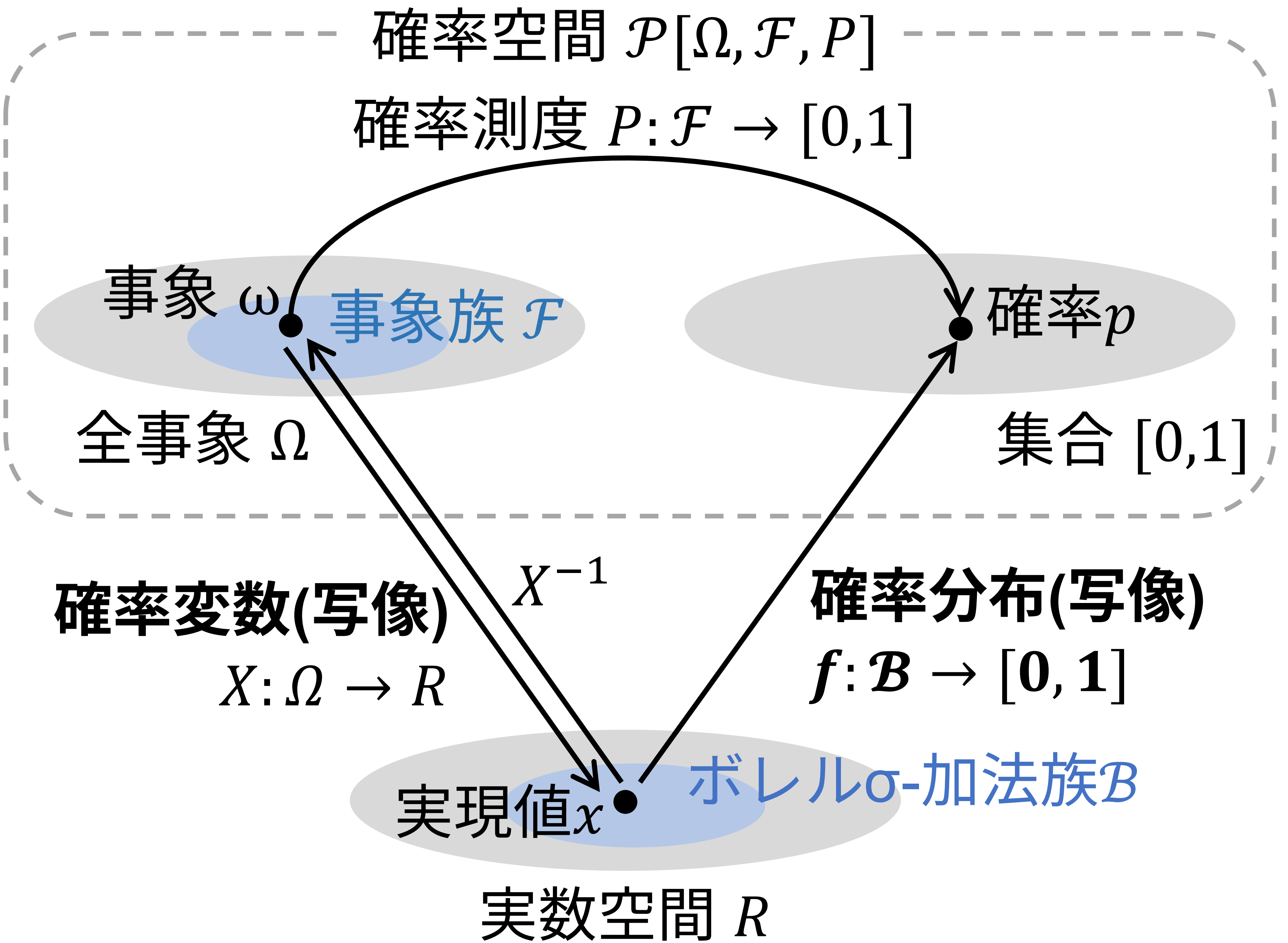
確率分布 は 写像



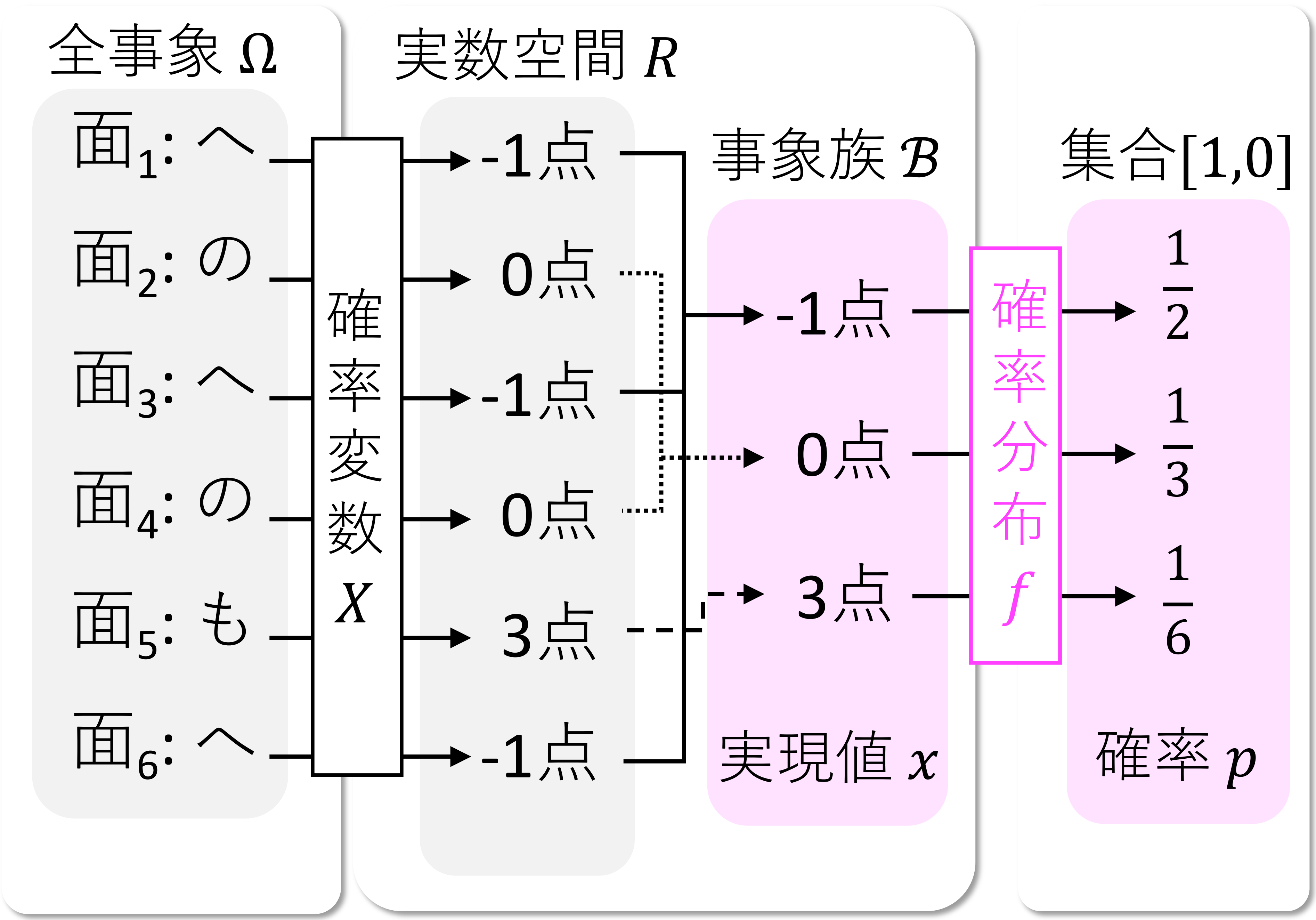
族$\mathcal{B}$が満たすべき<u>譲れない５条件</u>

$\rightarrow$ **ボレルσ-加法族**$\mathcal{B}$と呼ぶ

実数集合上で定義されるσ-加法族をボレルσ-加法族と呼んで差し支えない。



確率空間 $\mathcal{P}[\Omega, \mathcal{F}, P]$
確率測度 $P\colon \mathcal{F} \to [0,1]$
事象 ω
事象族 $\mathcal{F}$
確率p
全事象 Ω
集合 [0,1]
X^{-1}
確率変数(写像)
$X\colon \Omega \to R$
確率分布(写像)
$\boldsymbol{f}\colon \boldsymbol{\mathcal{B}} \to [\mathbf{0}, \mathbf{1}]$
ボレルσ-加法族$\mathcal{B}$
実現値x
実数空間 R



全事象 Ω
面$_1$: へ
面$_2$: の
面$_3$: へ
面$_4$: の
面$_5$: も
面$_6$: へ
確率変数 X
実数空間 R
-1点
0点
-1点
0点
3点
-1点
事象族 $\mathcal{B}$
-1点
0点
3点
実現値 x
確率分布 f
集合[1,0]
$\frac{1}{2}$
$\frac{1}{3}$
$\frac{1}{6}$
確率 p

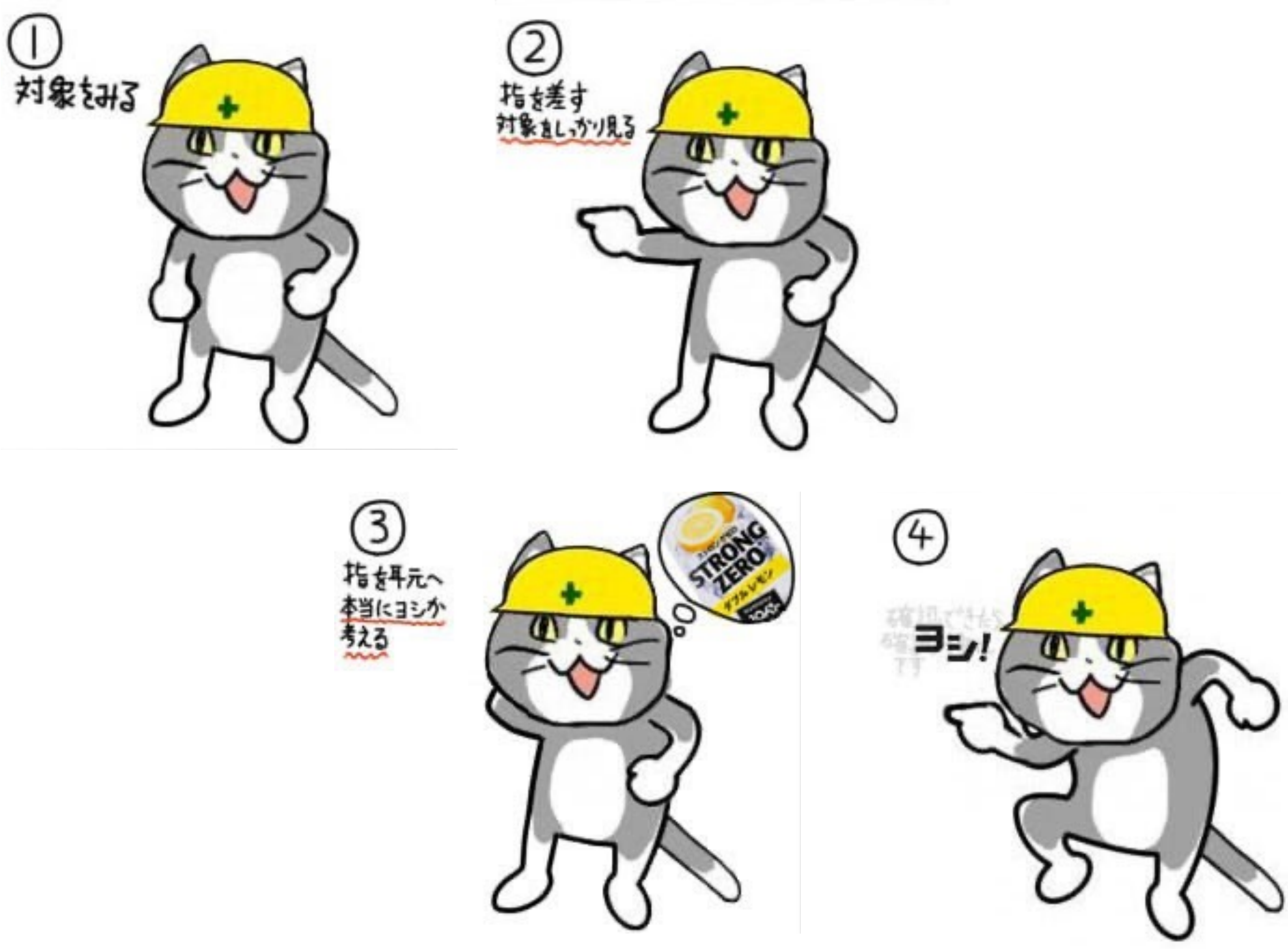
現在位置の確認

1. **写像**という概念を導入しました
2. **確率測度**Pを定義しました
3. **確率空間**$\mathcal{P}$を定義しました
4. **実数空間**R（**実現値**x）を導入しました
5. 実数空間Rと確率空間$\mathcal{P}$を写像で結びました

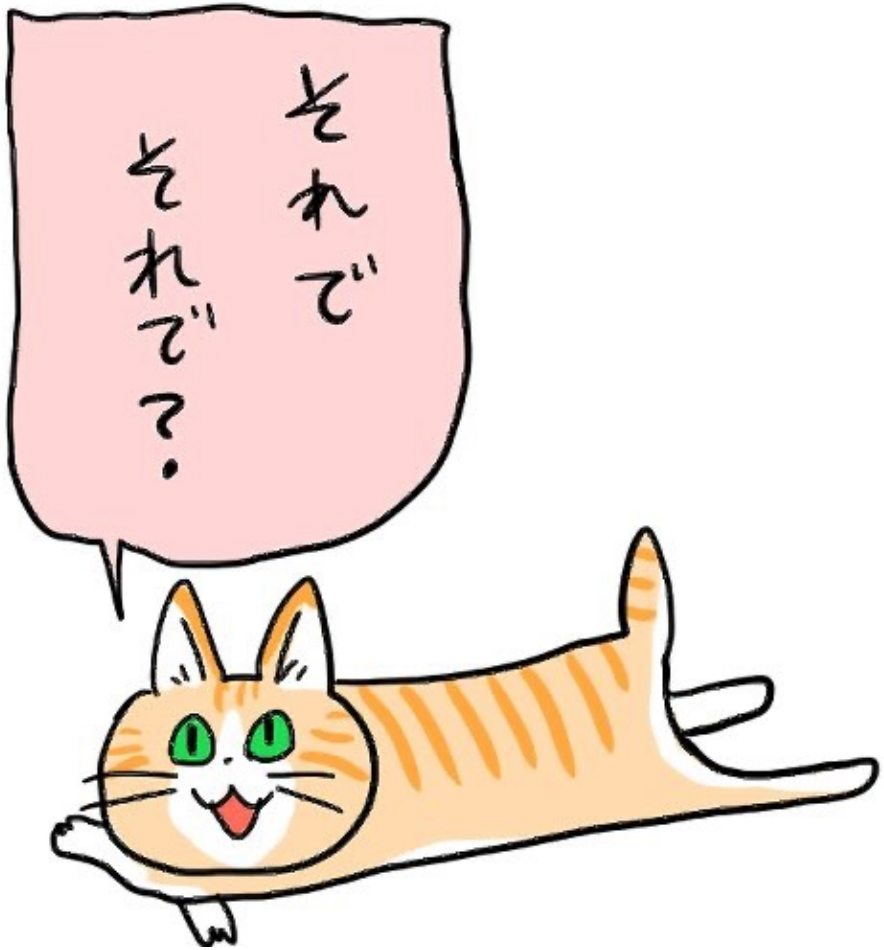
確率変数X ：事象ωから実現値xへの写像
確率分布f ：実現値xから確率pへの写像

これにより「実現値xに対する確率p 」を考えることが出来るようになりました。

（ヨシッ！）

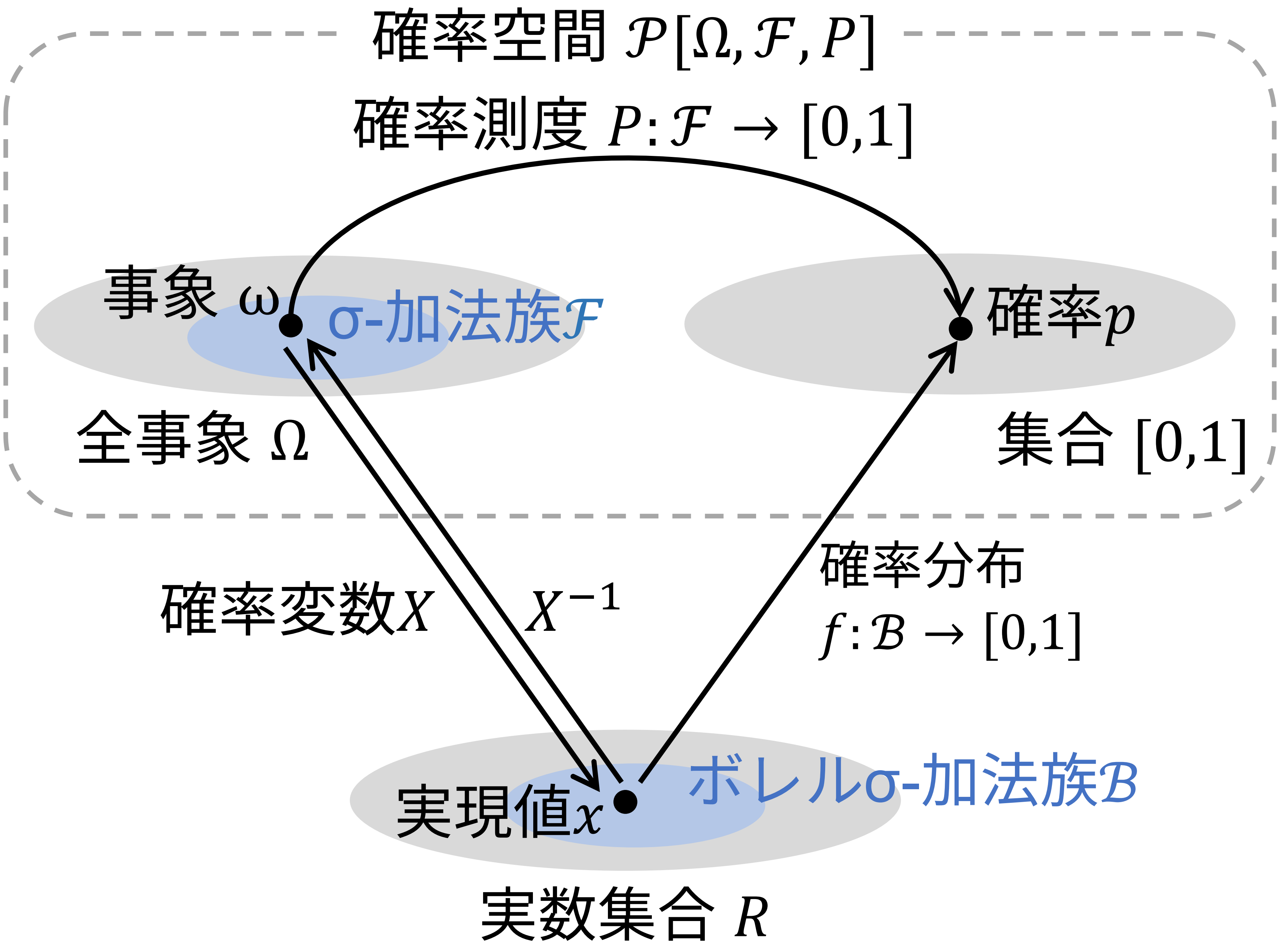


（https://i.imgur.com/J5yHe6L.jpg, オリジナルのリンク不明、ヨシ！じゃない、ごめんなさい）

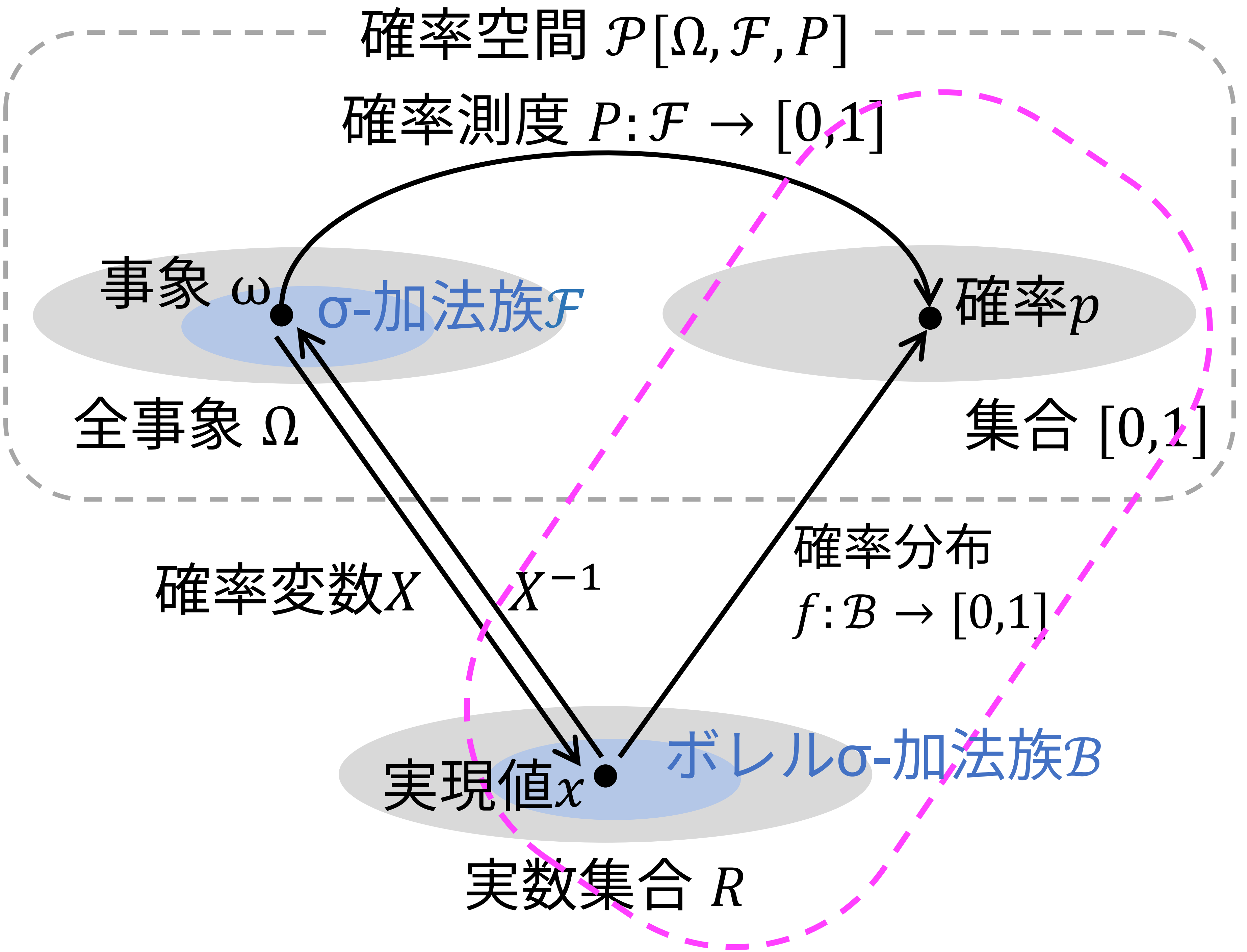


くまみね @kumamine

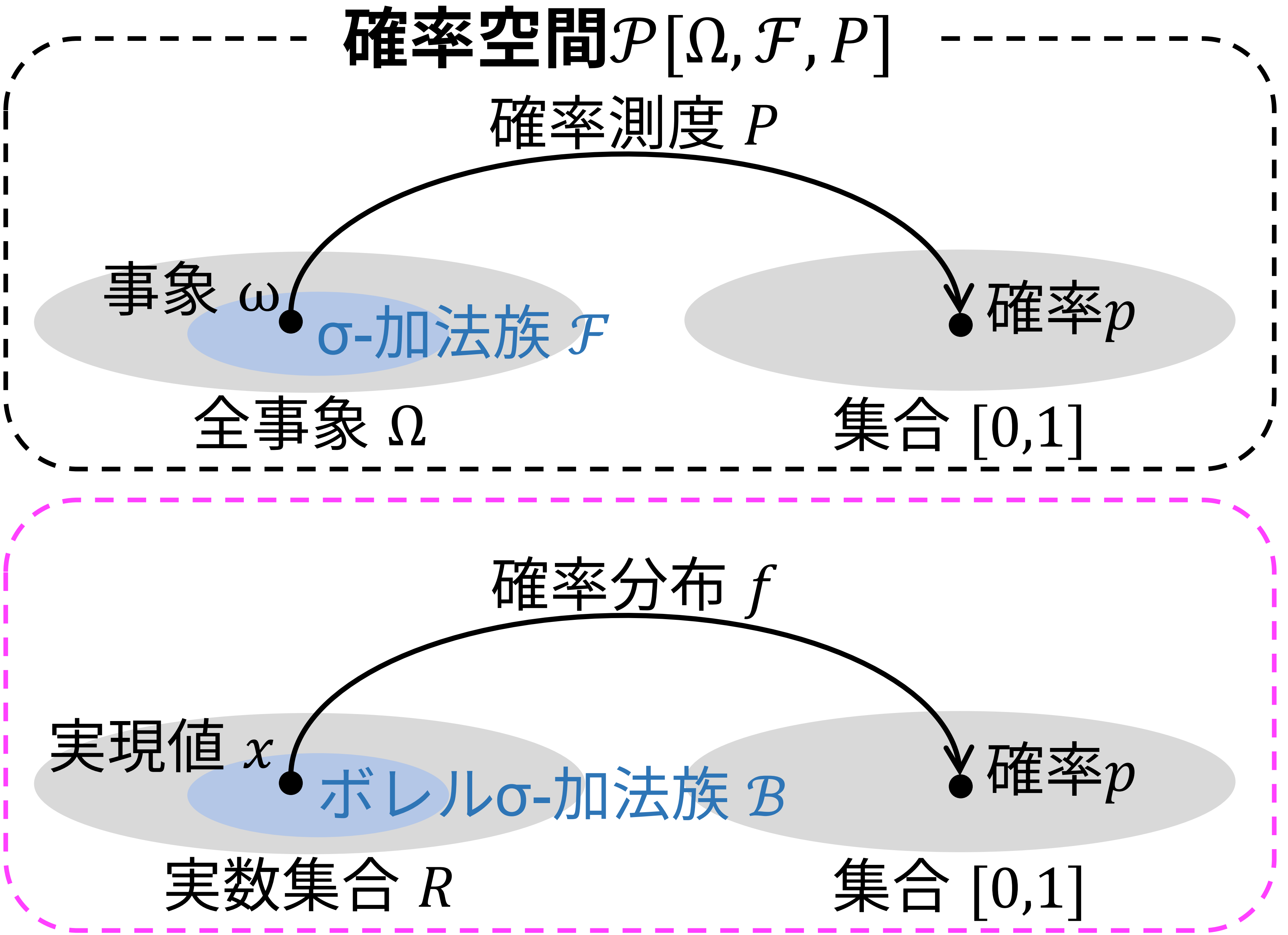
https://twitter.com/kumamine/status/956852321970831360?s=20&t=bXq3aqLCBlzaH_HR2pTPrg



確率空間 $\mathcal{P}[\Omega, \mathcal{F}, P]$
確率測度 $P: \mathcal{F} \rightarrow [0,1]$
事象 ω
σ-加法族$\mathcal{F}$
確率p
全事象 Ω
集合 [0,1]
確率分布
$f: \mathcal{B} \rightarrow [0,1]$
確率変数X
X^{-1}
ボレルσ-加法族$\mathcal{B}$
実現値x
実数集合 R



確率空間 $\mathcal{P}[\Omega, \mathcal{F}, P]$
確率測度 $P: \mathcal{F} \to [0,1]$
事象 ω
σ-加法族$\mathcal{F}$
確率p
全事象 Ω
集合 [0,1]
確率変数X
X^{-1}
確率分布
$f: \mathcal{B} \to [0,1]$
実現値x
ボレルσ-加法族$\mathcal{B}$
実数集合 R



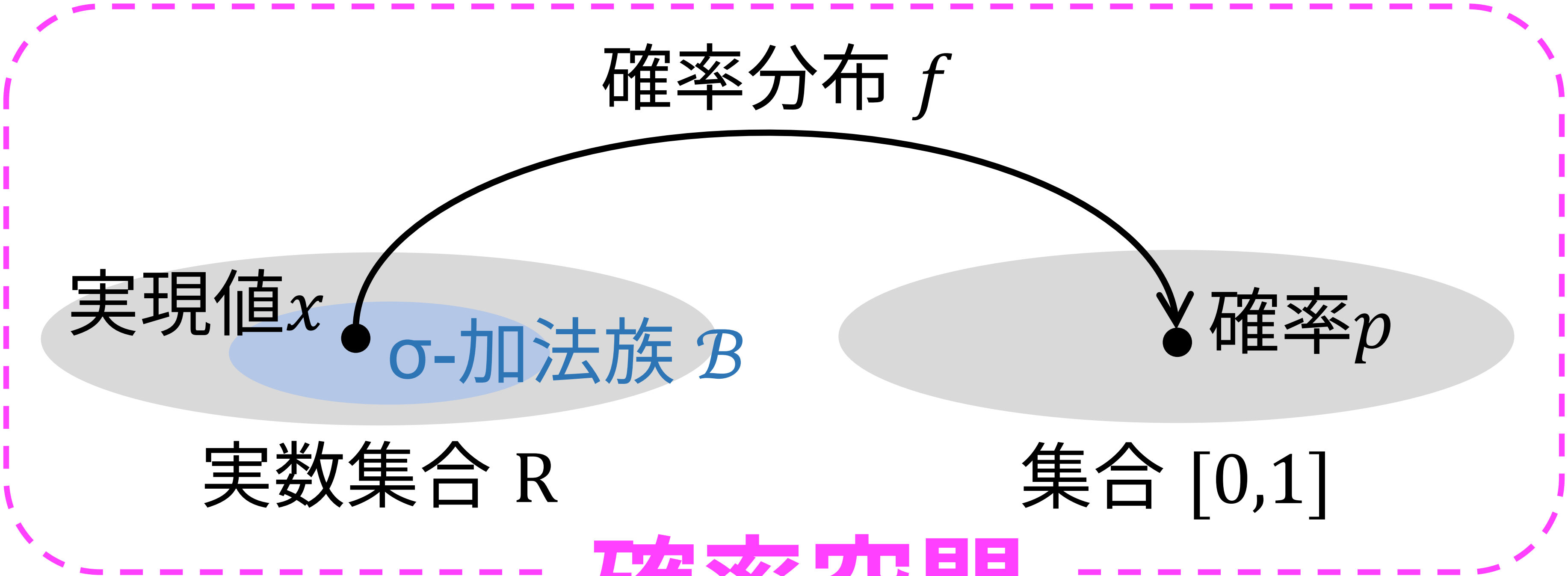
確率空間$\mathcal{P}[\Omega, \mathcal{F}, P]$
確率測度 P
事象 ω
σ-加法族 $\mathcal{F}$
確率p
全事象 Ω
集合 [0,1]
確率分布 f
実現値 x
ボレルσ-加法族 $\mathcal{B}$
確率p
実数集合 R
集合 [0,1]

お前はもう、

（ご唱和ください）

お前はもう、

確率空間だ。



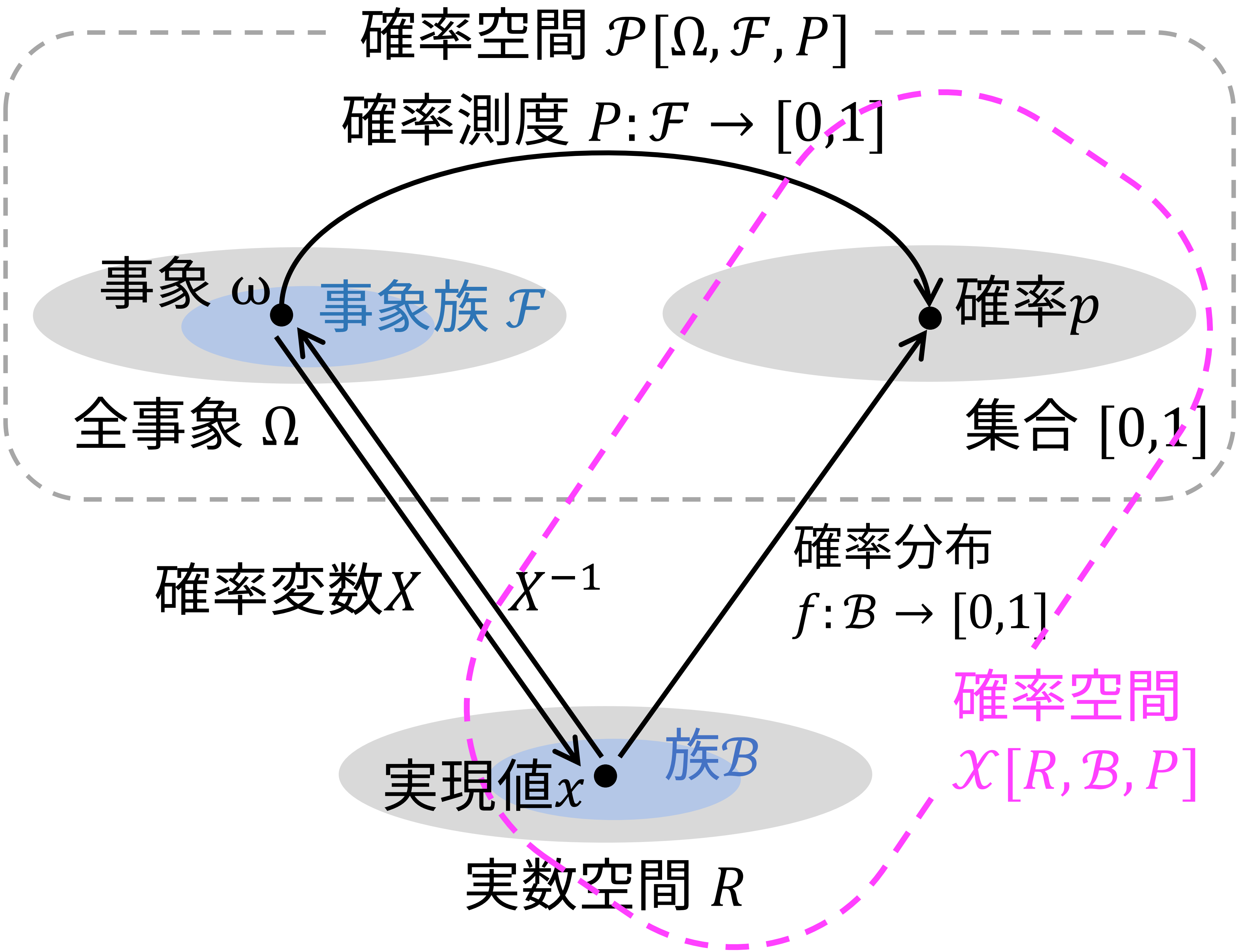
確率空間

$$\mathcal{X}[R,\mathcal{B},f]$$

全事象R: 起こりうる全ての事象を網羅した集合

事象族$\mathcal{B}$: σ-加法性を満たし写像のchannelを規定する

確率分布f: 実現値xに対し確率pを対応づける写像


確率空間 $\mathcal{P}[\Omega, \mathcal{F}, P]$
確率測度 $P: \mathcal{F} \to [0,1]$
事象 ω
事象族 $\mathcal{F}$
確率p
全事象 Ω
集合 [0,1]
確率分布
$f: \mathcal{B} \to [0,1]$
確率変数X
X^{-1}
確率空間
$\mathcal{X}[R, \mathcal{B}, P]$
族$\mathcal{B}$
実現値x
実数空間 R

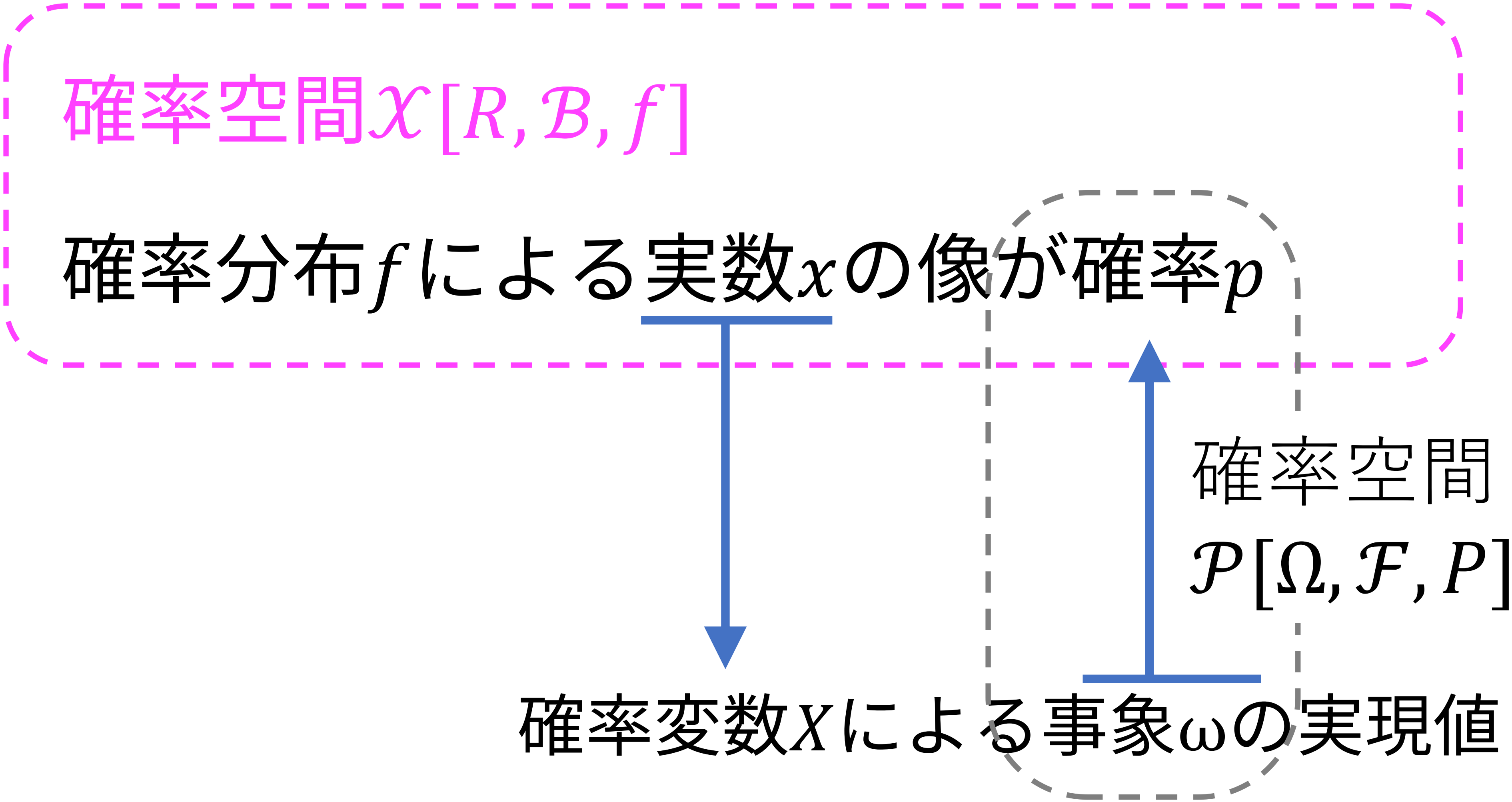
確率空間$\mathcal{X}[R,\mathcal{B},f]$

確率分布fによる実数xの像が確率p

確率空間$\mathcal{X}[R, \mathcal{B}, f]$

確率分布fによる実数xの像が確率p

確率変数Xによる事象ωの実現値


確率空間$\mathcal{X}[R,\mathcal{B},f]$
確率分布fによる実数xの像が確率p
確率空間
$\mathcal{P}[\Omega,\mathcal{F},P]$
確率変数Xによる事象ωの実現値

事象ωの確率変数Xによる実現値xは確率分布fに従って確率pに対応づけられる

事象ωの確率変数Xによる実現値xは確率分布fに従って確率pに対応づけられる

↓

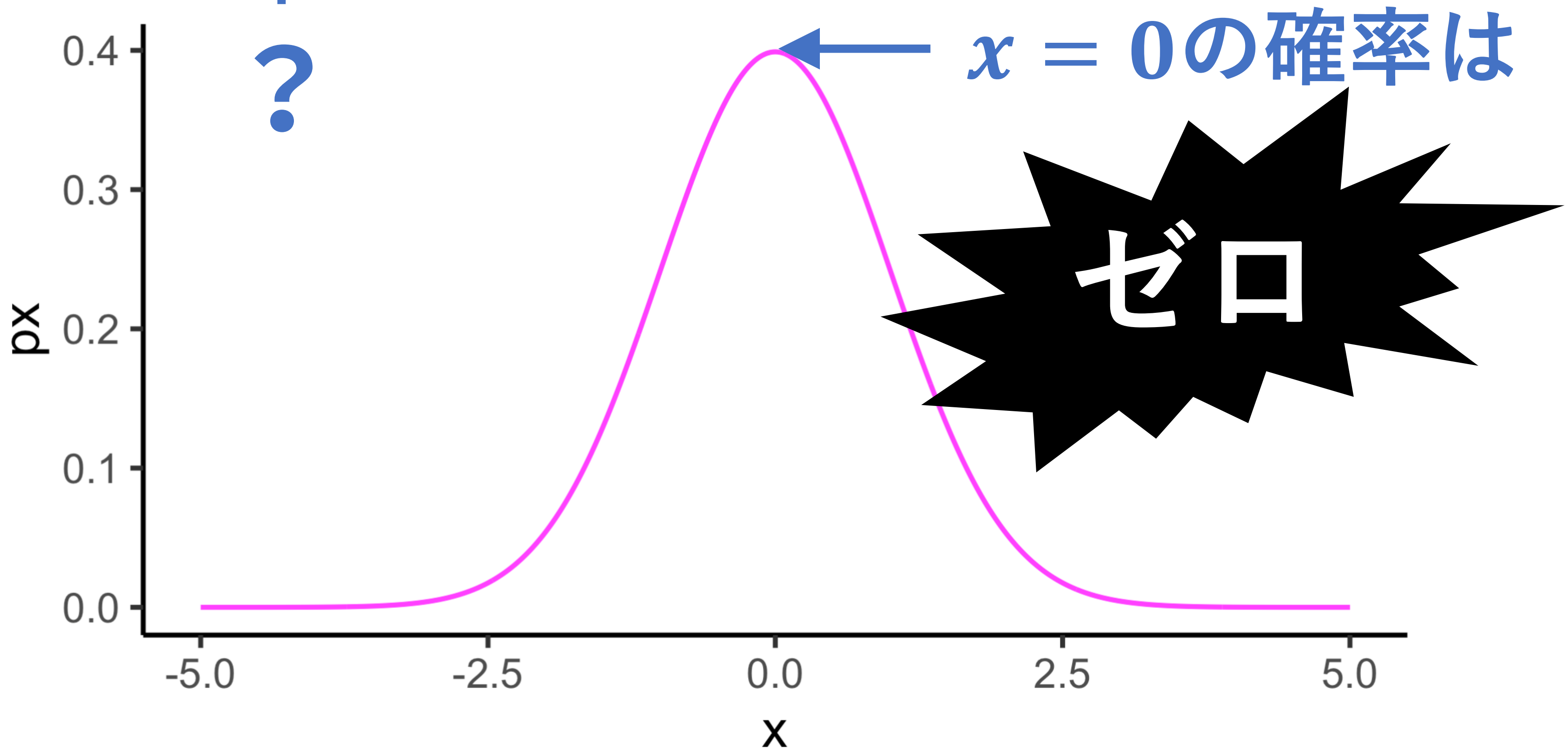
事象ωの**確率変数X**による実現値x**は確率分布fに従**って確率pに対応づけられる
う

$$X \sim N(0, 1)$$

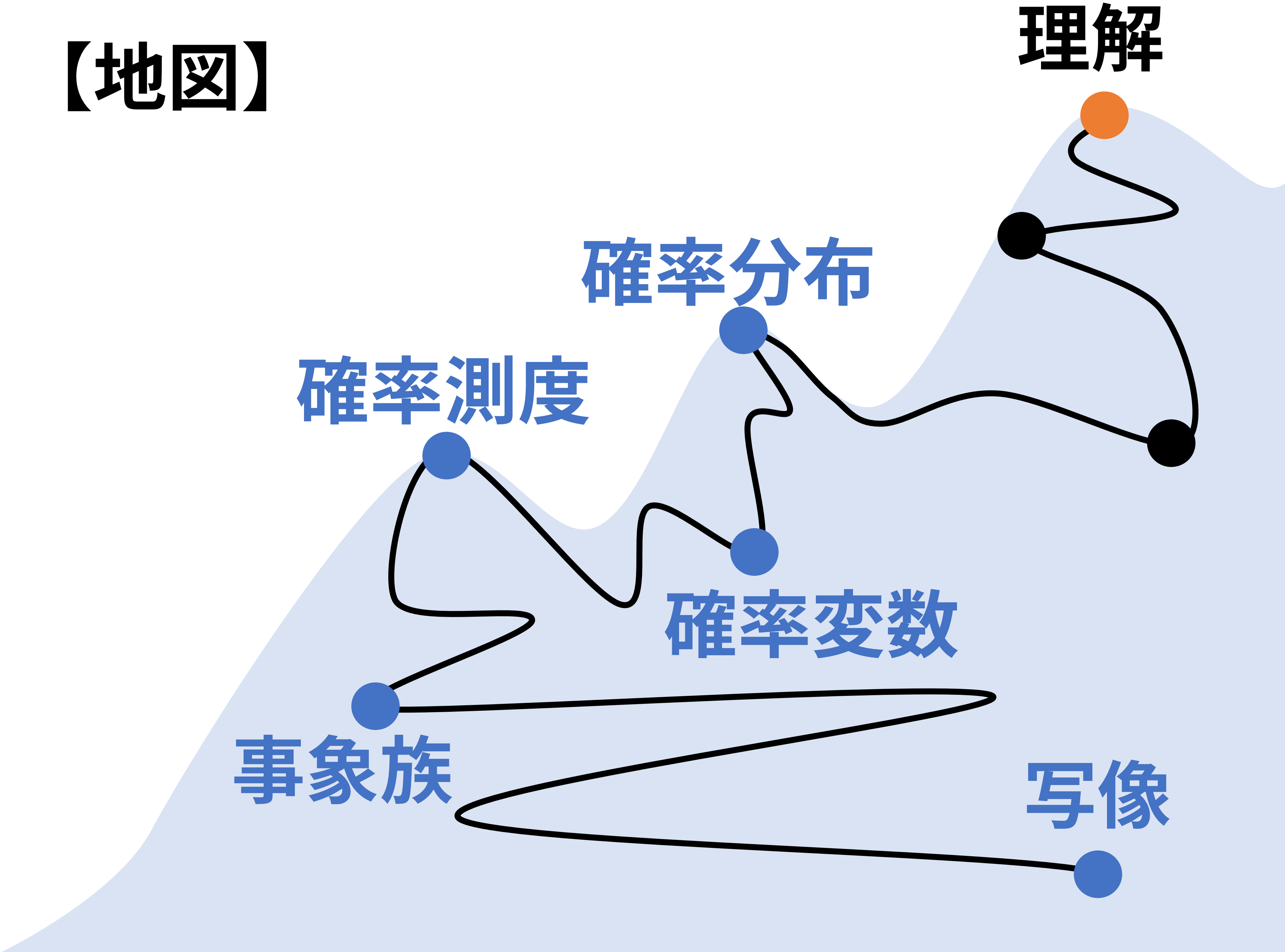
「**確率変数**X は**正規分布**$N(0,1)$に従う」

事象ωの確率変数Xによる実現値xは正規分布Nに従って確率pに対応づけられる

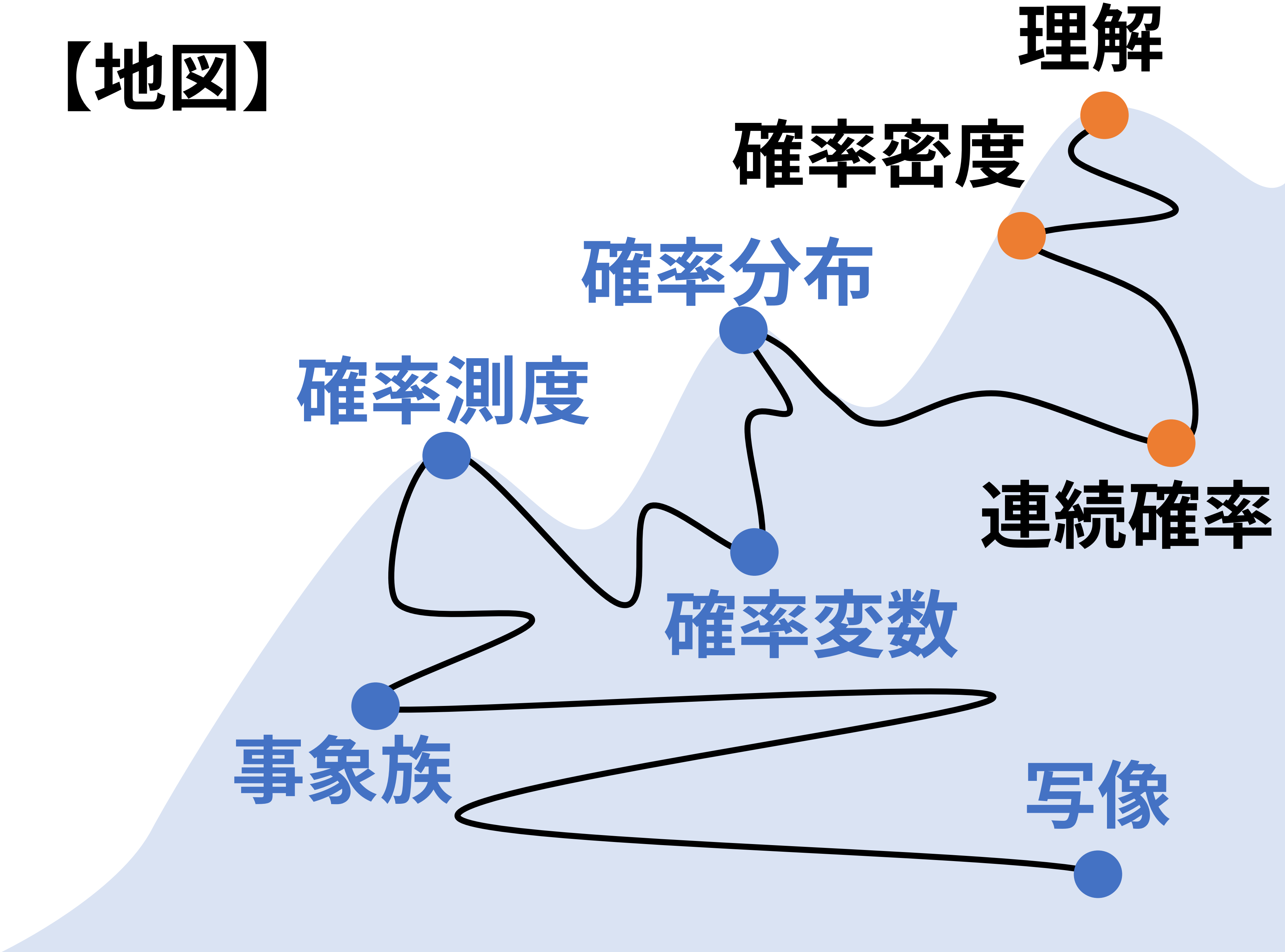
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

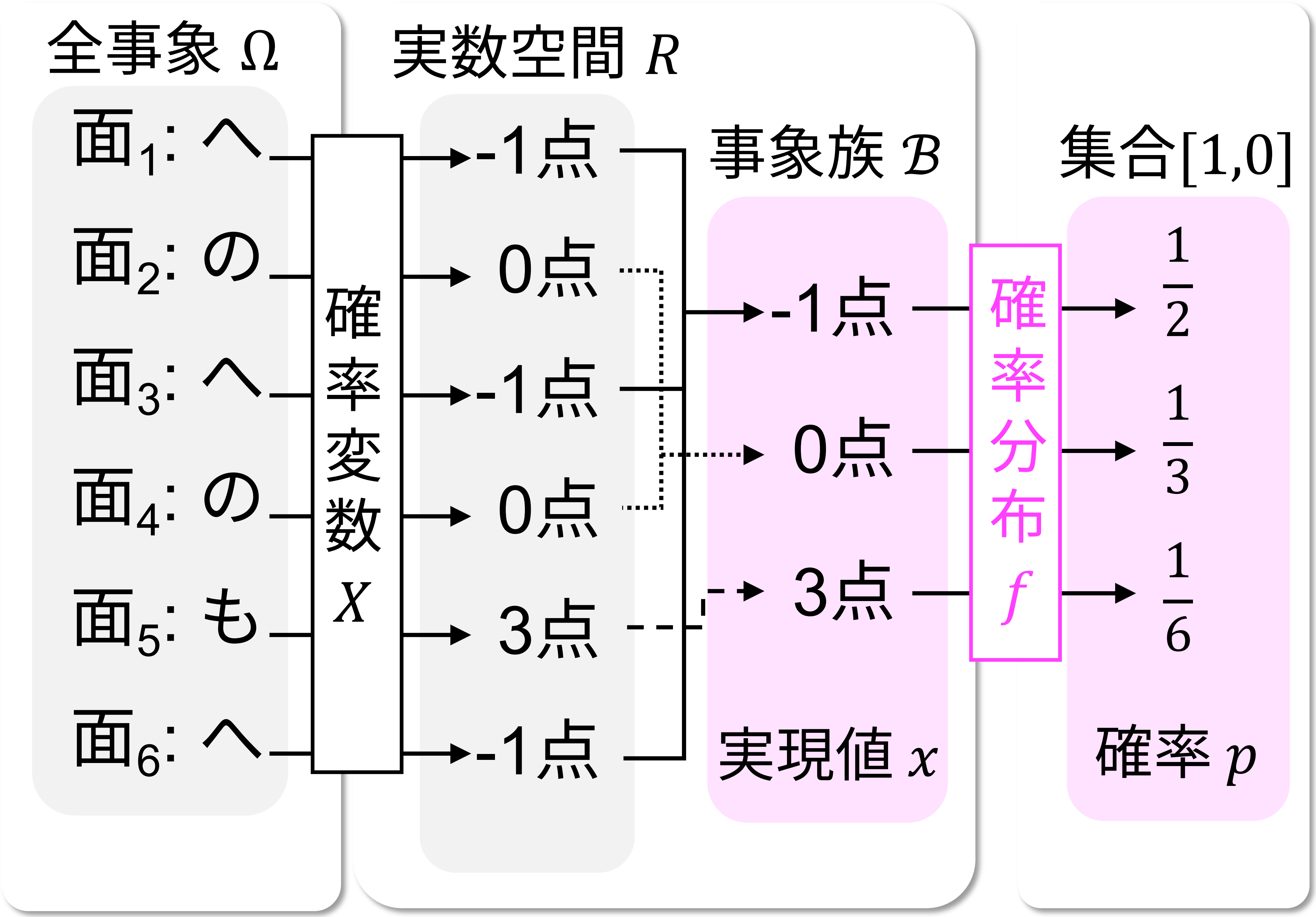


【地図】



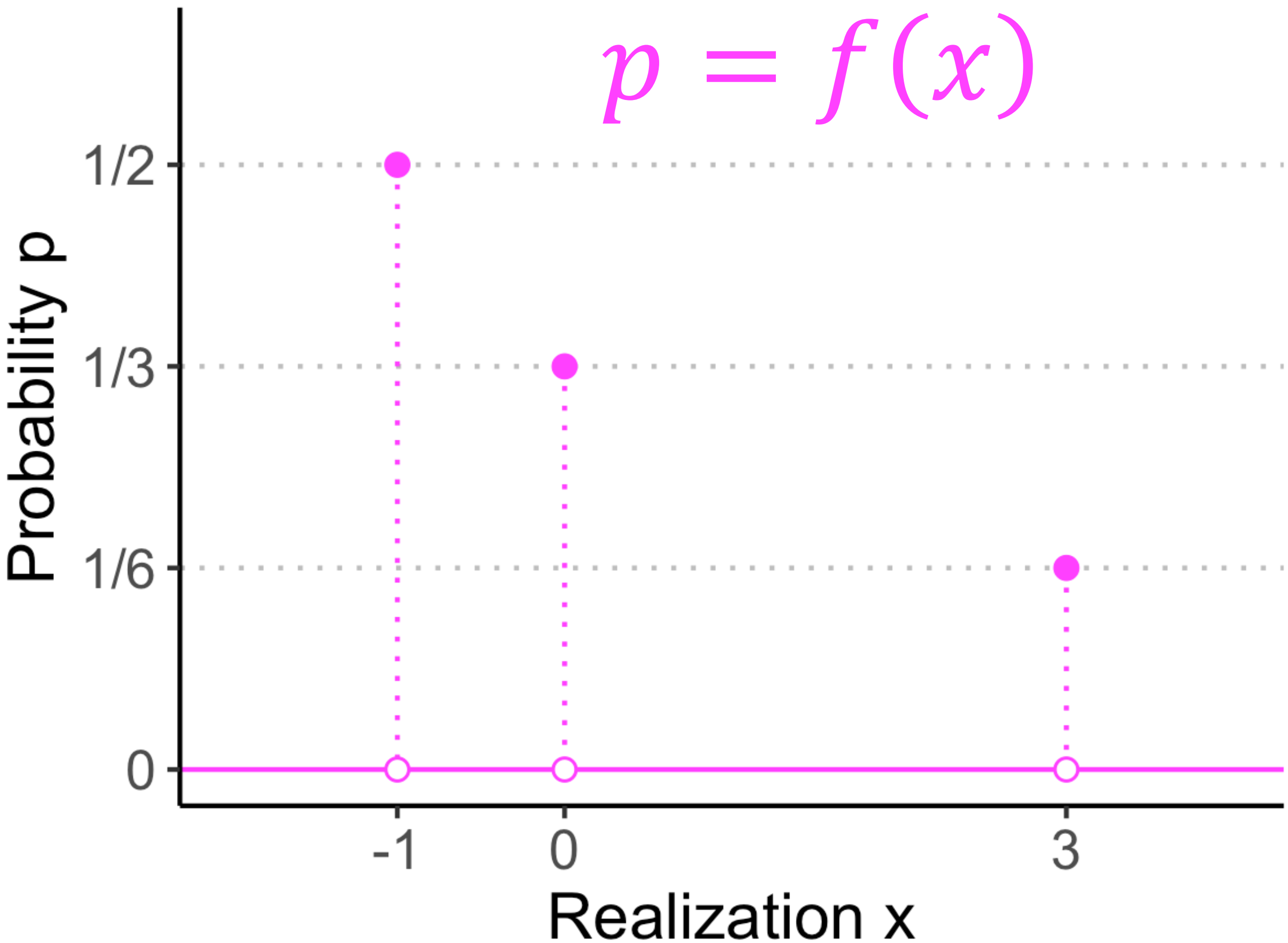
【地図】





全事象 Ω
面$_1$: へ
面$_2$: の
面$_3$: へ
面$_4$: の
面$_5$: も
面$_6$: へ
確率変数 X
実数空間 R
-1点
0点
-1点
0点
3点
-1点
事象族 $\mathcal{B}$
-1点
0点
3点
実現値 x
確率分布 f
集合[1,0]
$\frac{1}{2}$
$\frac{1}{3}$
$\frac{1}{6}$
確率 p

確率分布fを可視化する



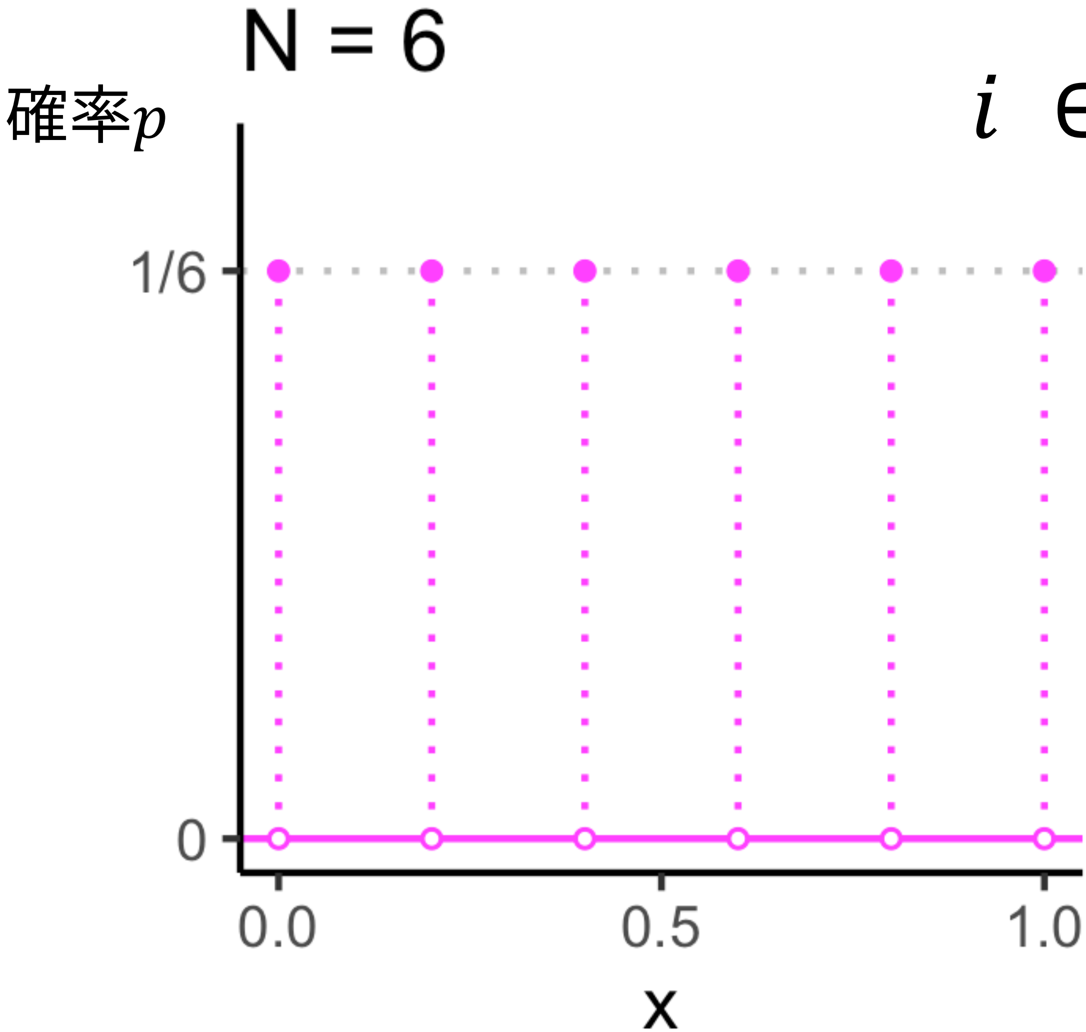
理想的なN面体サイコロ

- 必ず1つの面が出る
- 面は均等に出る
- 各面には数値(実現値x)が1つ書かれている
- x ∈ seq(form = 0, to = 1, length = N)
- 確率分布$f: x \longmapsto p$すなわち$f(x) = p$とする

理想的なN面体サイコロ

$$p_i = f(x_i) = \frac{1}{6}$$

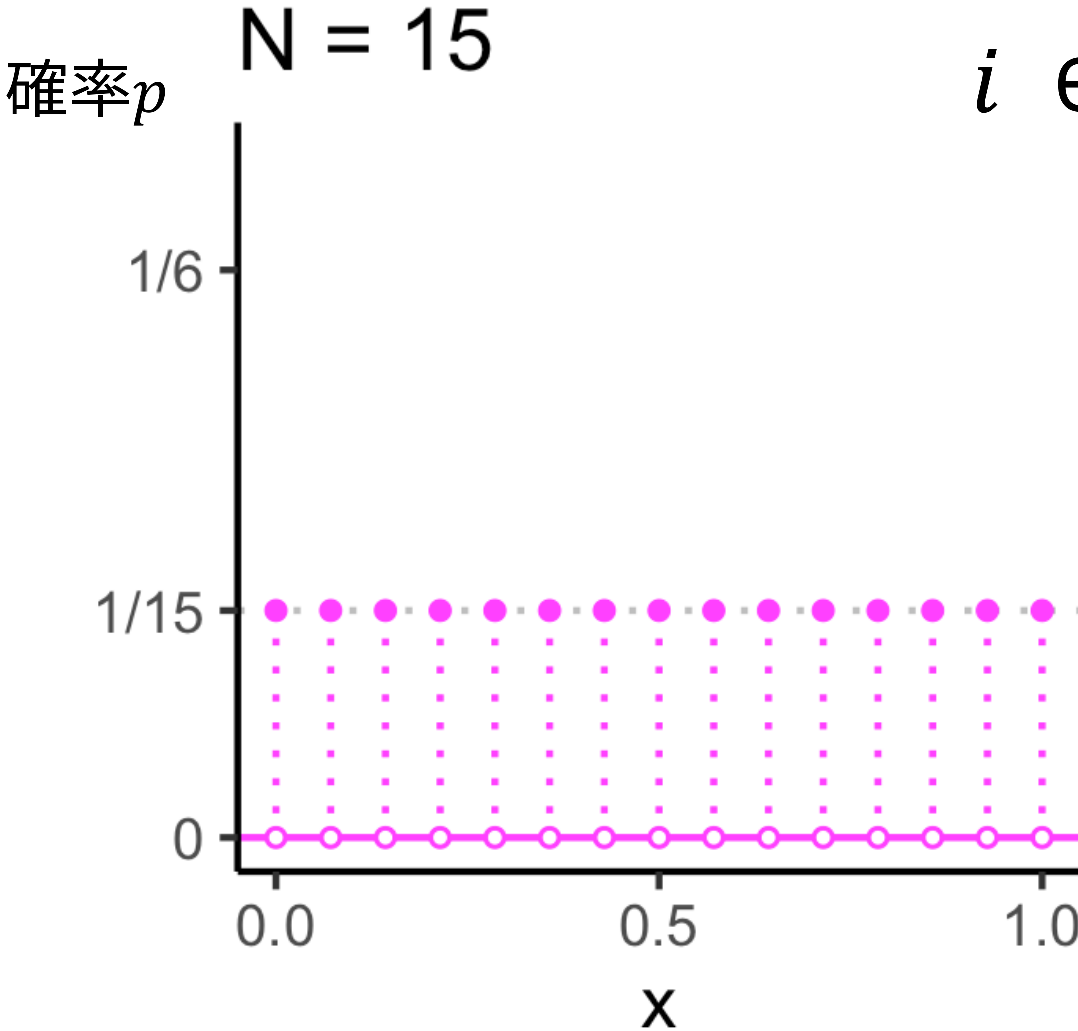
$$i \in \{1, 2, \ldots, 6\}$$



理想的なN面体サイコロ

$$p_i = f(x_i) = \frac{1}{15}$$

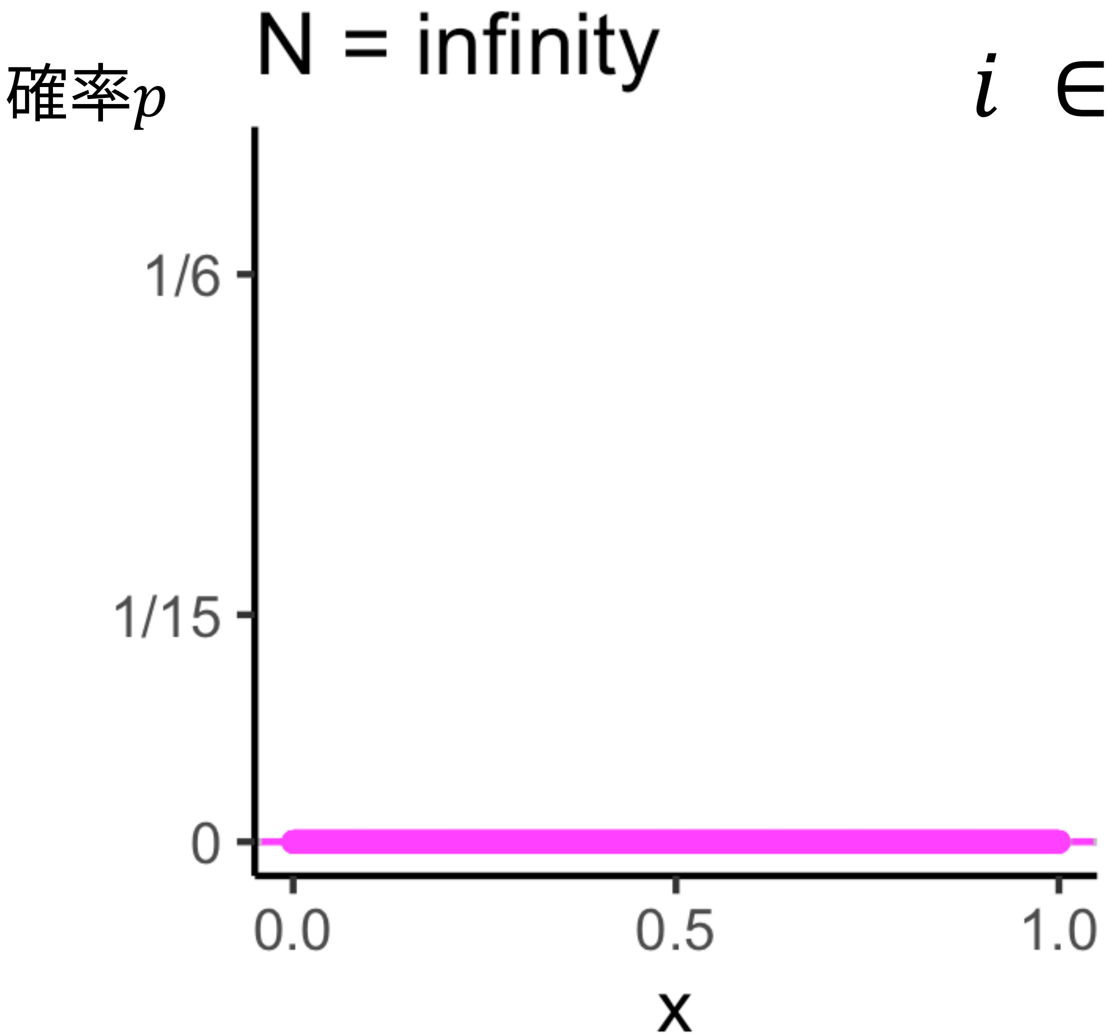
$$i \in \{1, 2, \ldots, 15\}$$



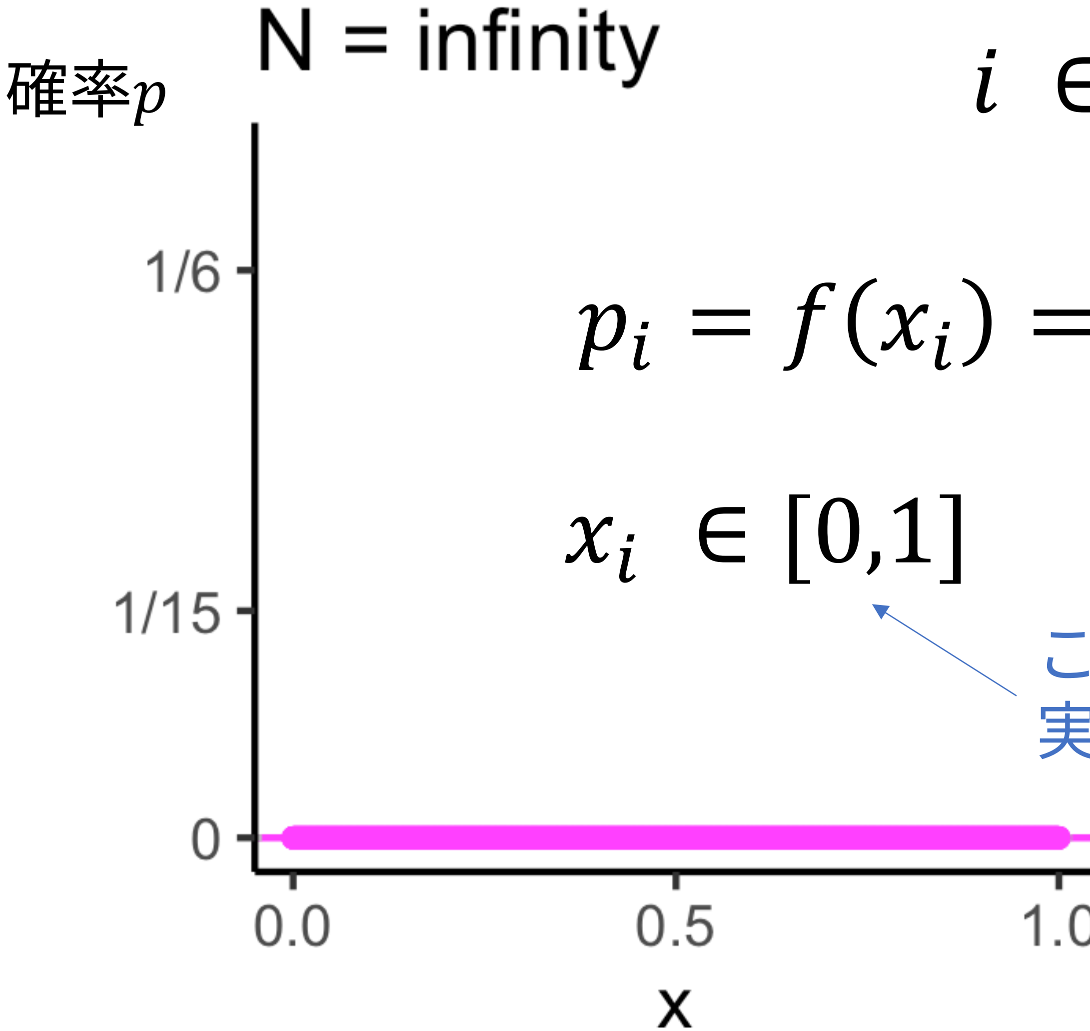
理想的なN面体サイコロ

$$p_i = f(x_i) = \frac{1}{\infty}$$

$$i \in \{1, 2, \dots, \infty\}$$



理想的なN面体サイコロ



$$i \in \{1, 2, \dots, \infty\}$$

$$p_i = f(x_i) = \frac{1}{\infty} = 0$$

$$x_i \in [0,1]$$

この区間に含まれる実数全体

連続確率分布f

実現値xに対し確率pを対応づける写像のうちxが連続した1つの区間$[a, b]$で定義されるもの。ただし$(a, b \in R, a < b)$を満たす。

連続確率分布f

実現値xに対し確率pを対応づける写像のうちxが連続した1つの区間$[a, b]$で定義されるもの。ただし$(a, b \in R, a < b)$を満たす。

特定の実現値x_iに対する確率p_iは必ず0になる。

$$p_i = f(x_i) = \frac{1}{\infty} = 0$$

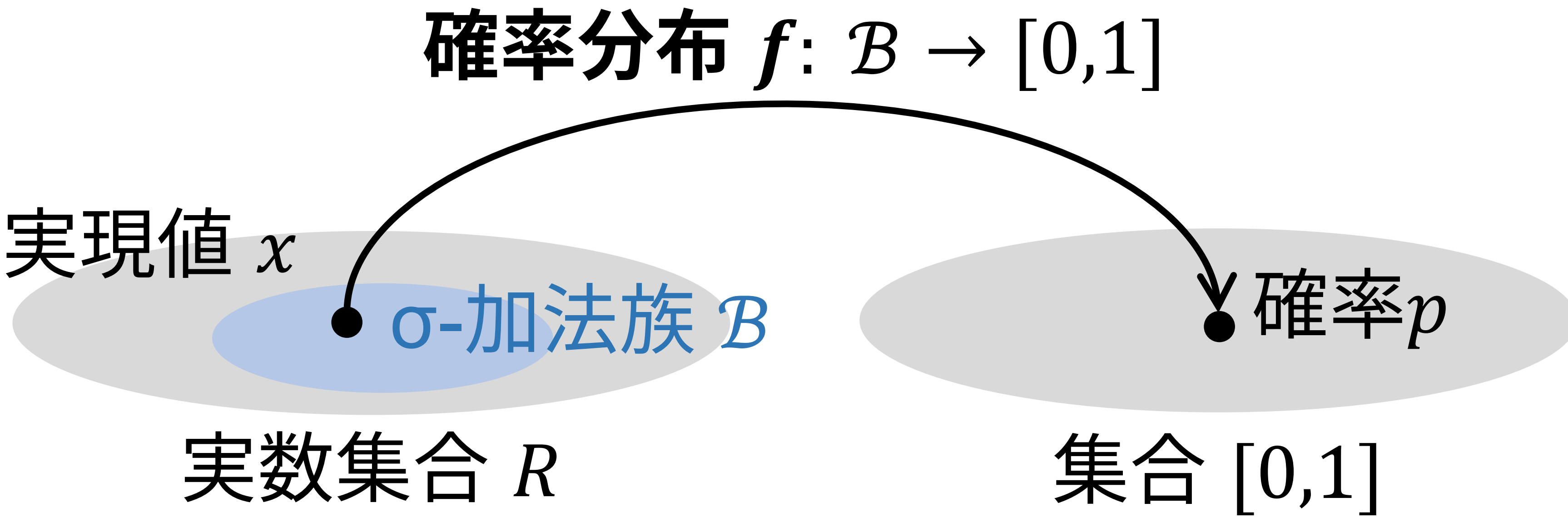
確率p_iは実現値x_iの「生じやすさ」を表す数値

$x \in [a, b]$だと全てのx_iについて$p_i = 0$

従って「どの面も出ない」

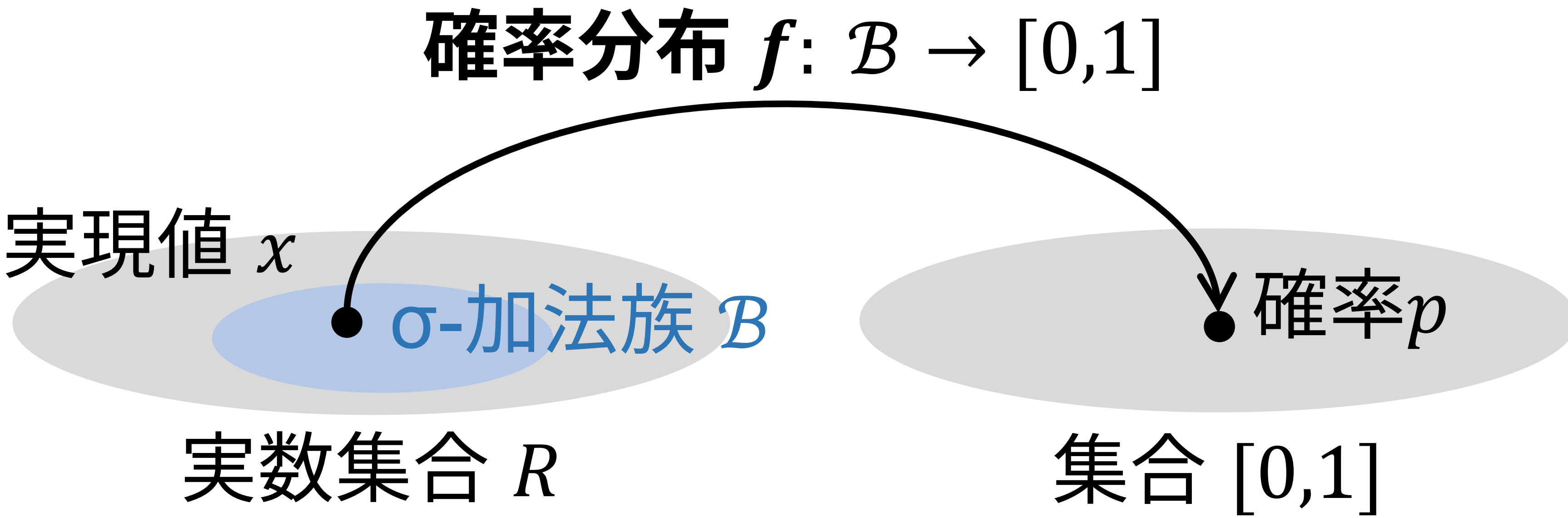
矛盾

「必ず1つの面が出る」（定義）



確率分布 fが満たすべき<u>譲れない３条件</u>

1. 写像f: $\mathcal{B} \to [0,1]$
2. 全事象の確率$f(R) = 1$
3. 可算加法性を満たす



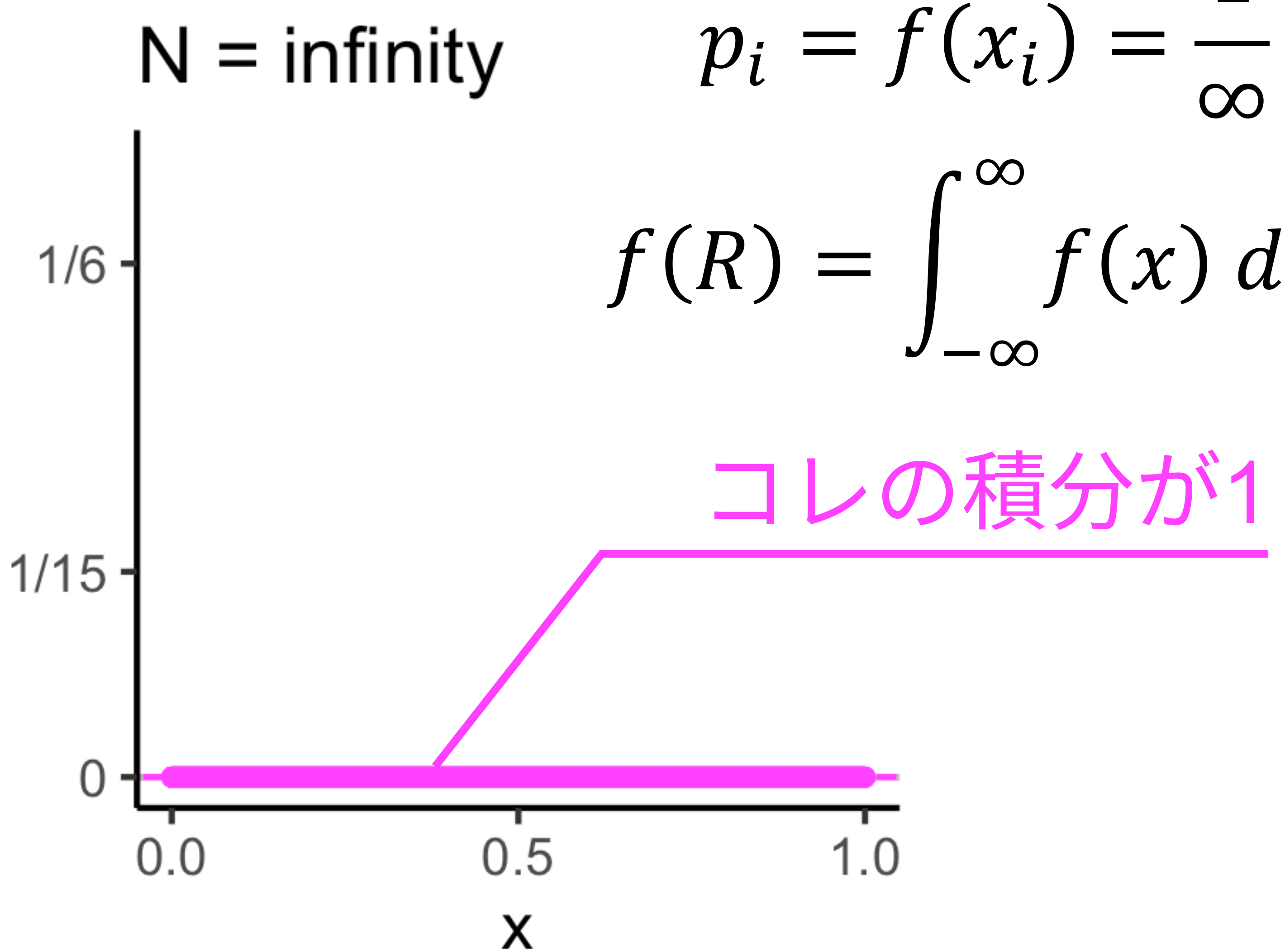
確率分布 fが満たすべき譲れない３条件より

$$f(R) = \int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} p\,dx = 1$$

理想的なN面体サイコロ

$$p_i = f(x_i) = \frac{1}{\infty} = 0$$

$$f(R) = \int_{-\infty}^{\infty} f(x)\,dx = 1$$



そうだ、**累積確率**を考えよう

累積確率F

連続確率分布fについて

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

離散確率分布fについて

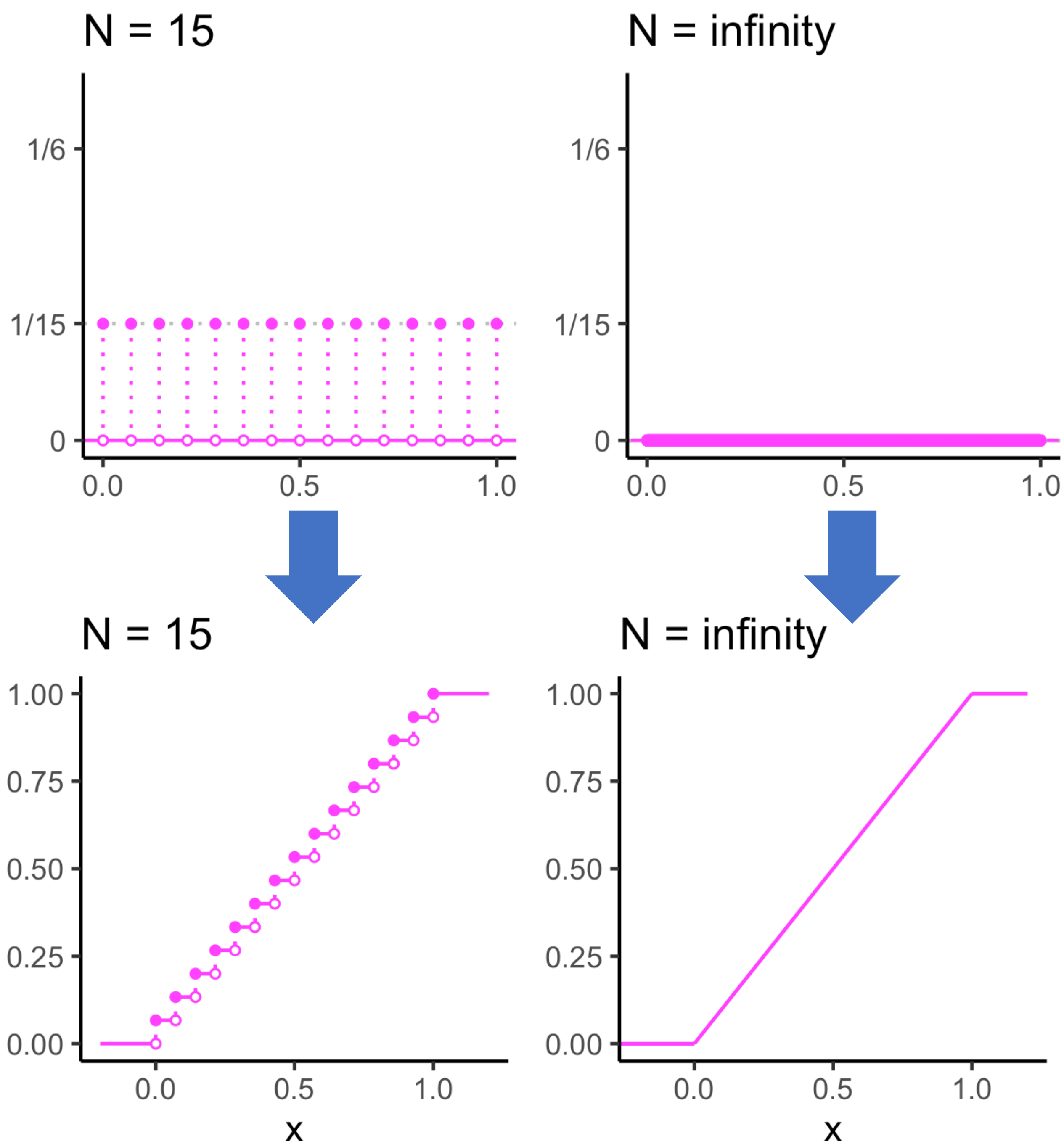
$$F(x) = \sum_{x_i \le x} f(x_i)$$

確率分布

$f(x)$

累積確率

$F(x)$



連続確率分布fについて

確率分布 $f(x) = \frac{1}{\infty}$

累積確率 $F(x) = \int_{-\infty}^{x} f(x)\, dx$

無限に小さい数を足し上げている

累積確率Fはxの定義域内で微分可能

微分可能？よろしい、
ならば**微分**しよう。

連続確率分布fについて

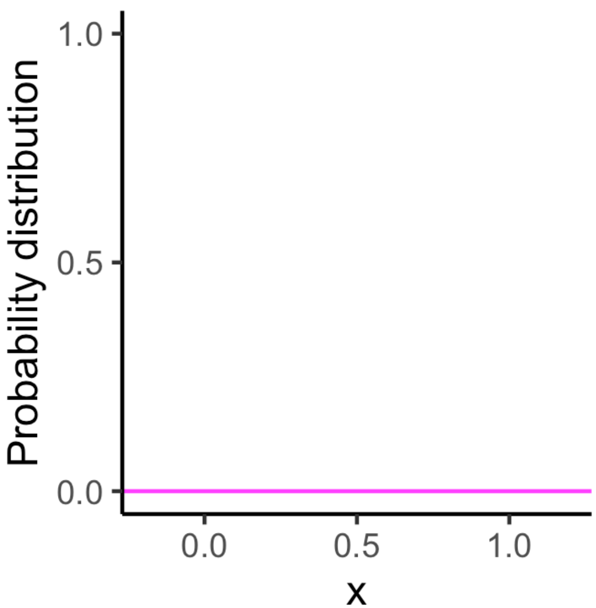
確率分布 $$f(x) = \frac{1}{\infty}$$

累積確率 $$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

確率密度 $$\mathrm{f}(x) = \frac{d}{dx} F(x)$$

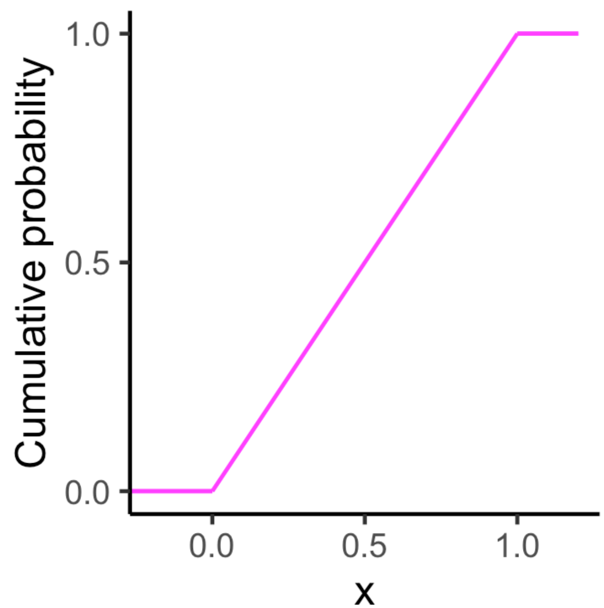
確率分布

$$f(x)$$



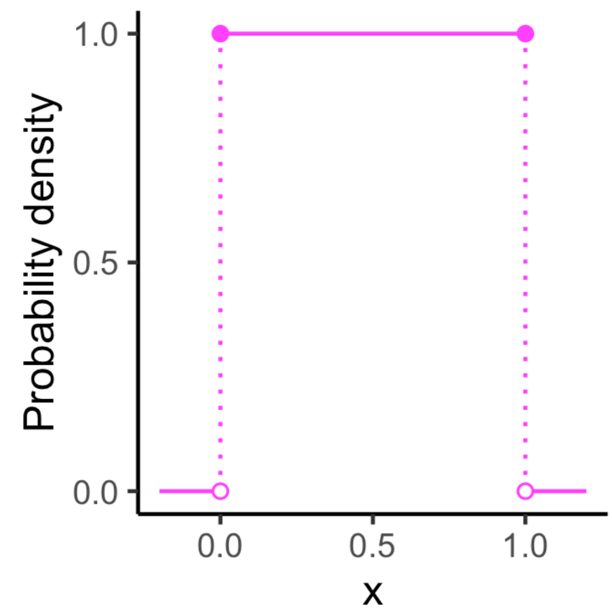
累積確率

$$F(x)$$

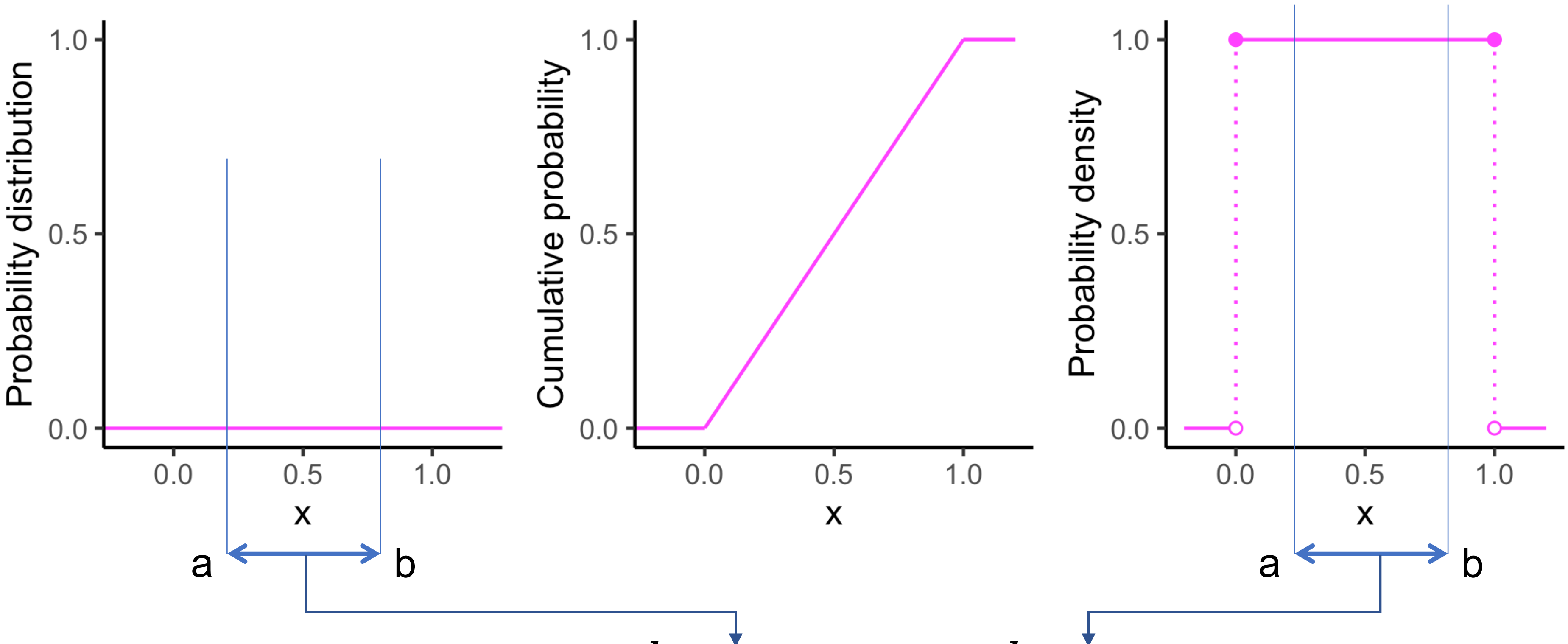


確率密度

$$\mathrm{f}(x)$$



確率分布 $f(x)$　　累積確率 $F(x)$　　確率密度 $\mathrm{f}(x)$



$$p(a \le x \le b) = \int_a^b f(x)dx = \int_a^b \mathrm{f}(x)dx$$

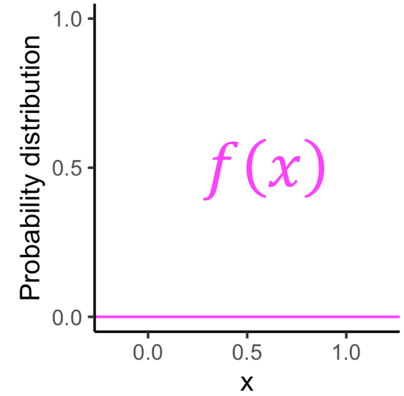
連続確率分布fについて

確率分布 $$f(x) = \frac{1}{\infty}$$

累積確率 $$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

確率密度 $$\mathrm{f}(x) = \frac{d}{dx}F(x)$$

例外なく定式化できない

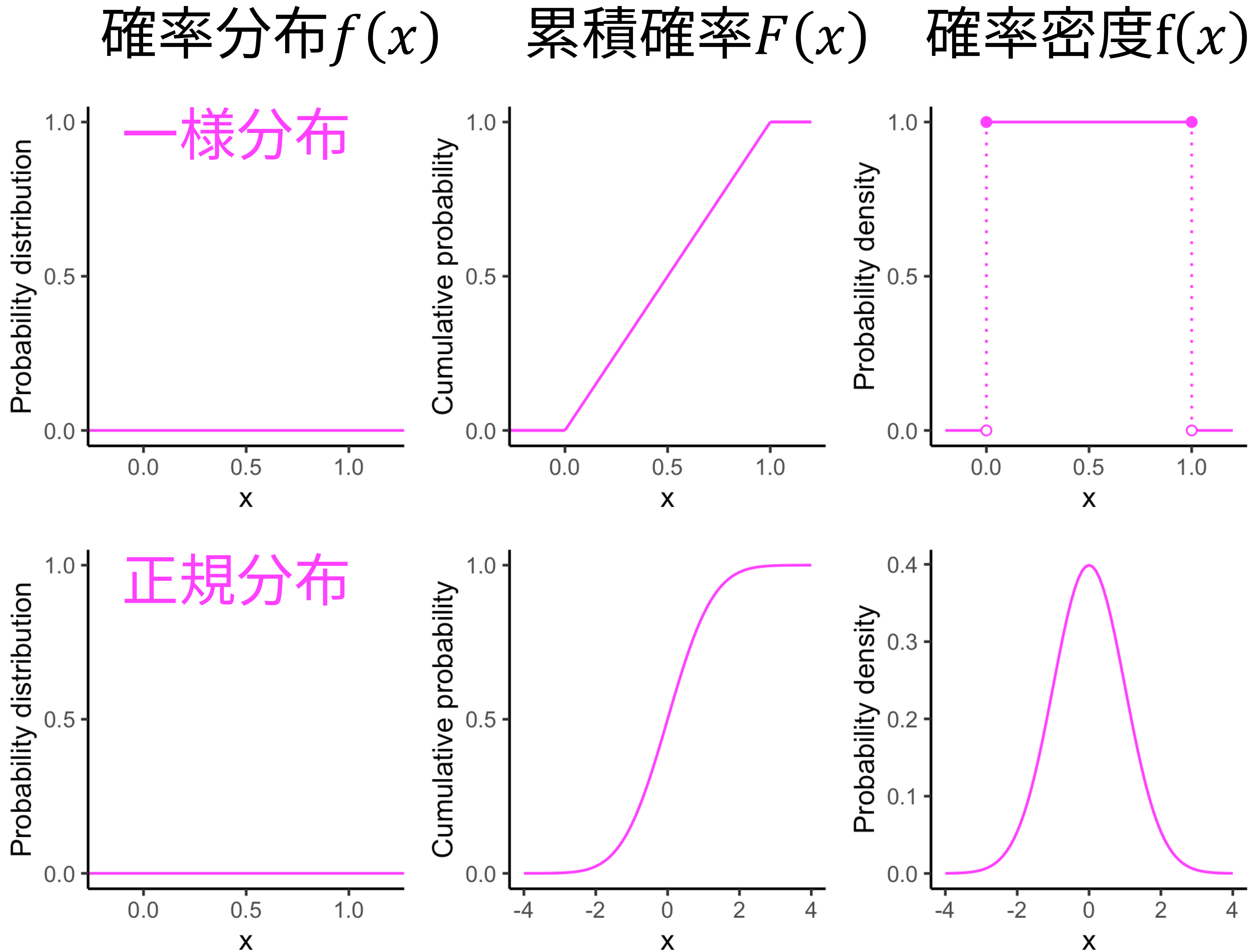


こっちで定式化する

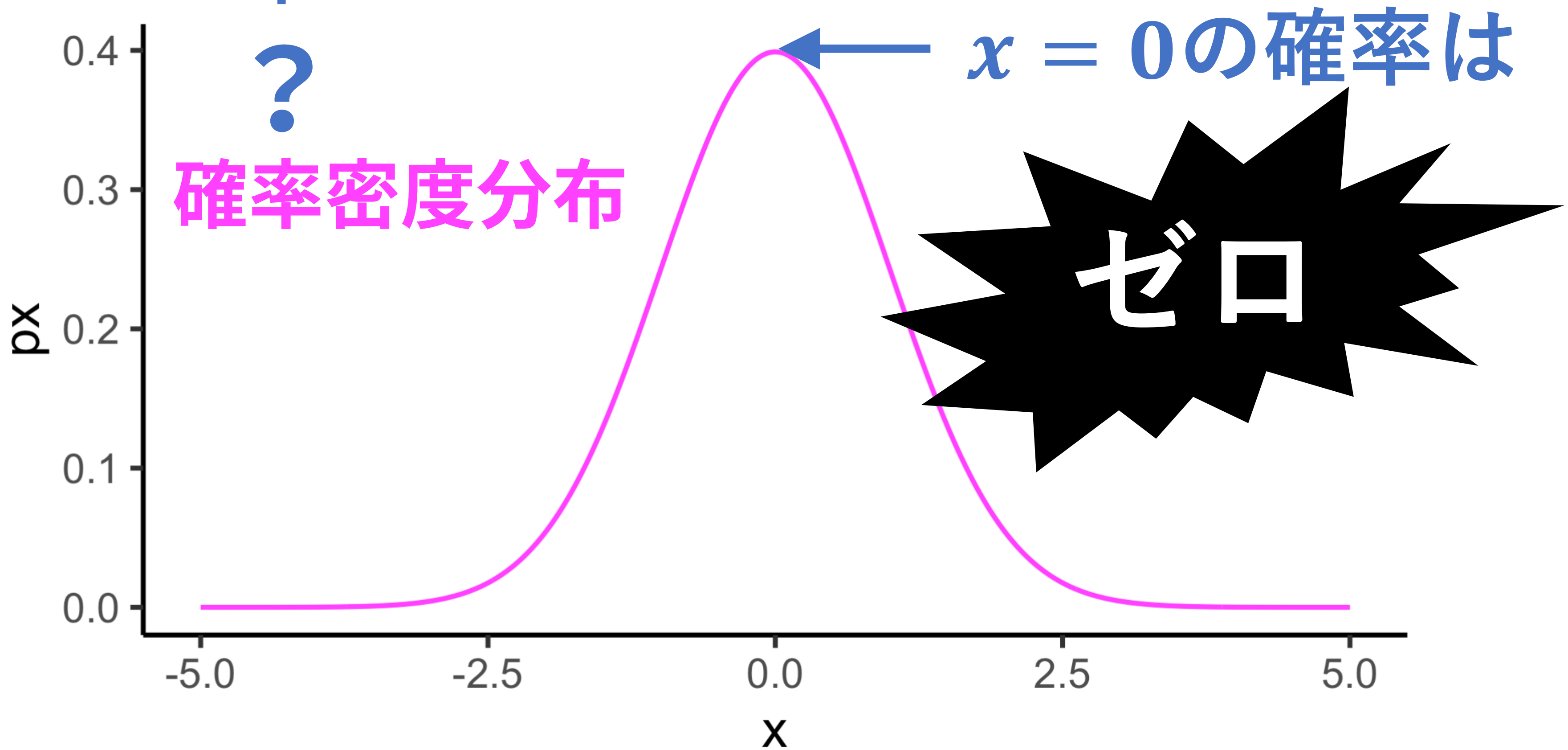
正規分布の**確率密度**f

$$\mathrm{f}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

例のあの式！


確率分布$f(x)$
累積確率$F(x)$
確率密度f(x)
一様分布
正規分布
Probability distribution
Cumulative probability
Probability density
x

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



【目標】

$$X \sim N(\mu = 0, \sigma = 1)$$

「**確率変数**X は**正規分布**$N(0,1)$に従う」

達成！！

呼び方いろいろ問題 (他にもあるかも)

$f(x)$	$F(x)$	$\mathrm{f}(x)$
確率分布関数	**累積確率**分布関数	**確率密度**分布関数
確率分布関数	累積確率分布関数	確率密度分布関数
確率分布関数	累積確率分布関数	確率密度分布関数
確率質量関数 (実現値が離散量の場合の呼び方)	累積確率分布関数	確率密度分布関数
	累積確率分布関数	

- 確率の話
  - 確率論の基礎
  - 確率分布 in R

確率分布を取り扱うための関数

	$f(x)$ 確率分布に従う乱数	$F(x)$ 累積確率	$F^{-1}(x)$ 確率点	$\mathrm{f}(x)$ 確率密度
	r*()**	**p***()**	**q***()**	**d***()**
正規分布	rnorm()	pnorm()	qnorm()	dnorm()
一様分布	runif()	punif()	qunif()	dunif()
f分布	rf()	pf()	qf()	df()
t分布	rt()	pt()	qt()	dt()
カイ²分布	rchisq()	pchisq()	qchisq()	dchisq()
⋮				

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

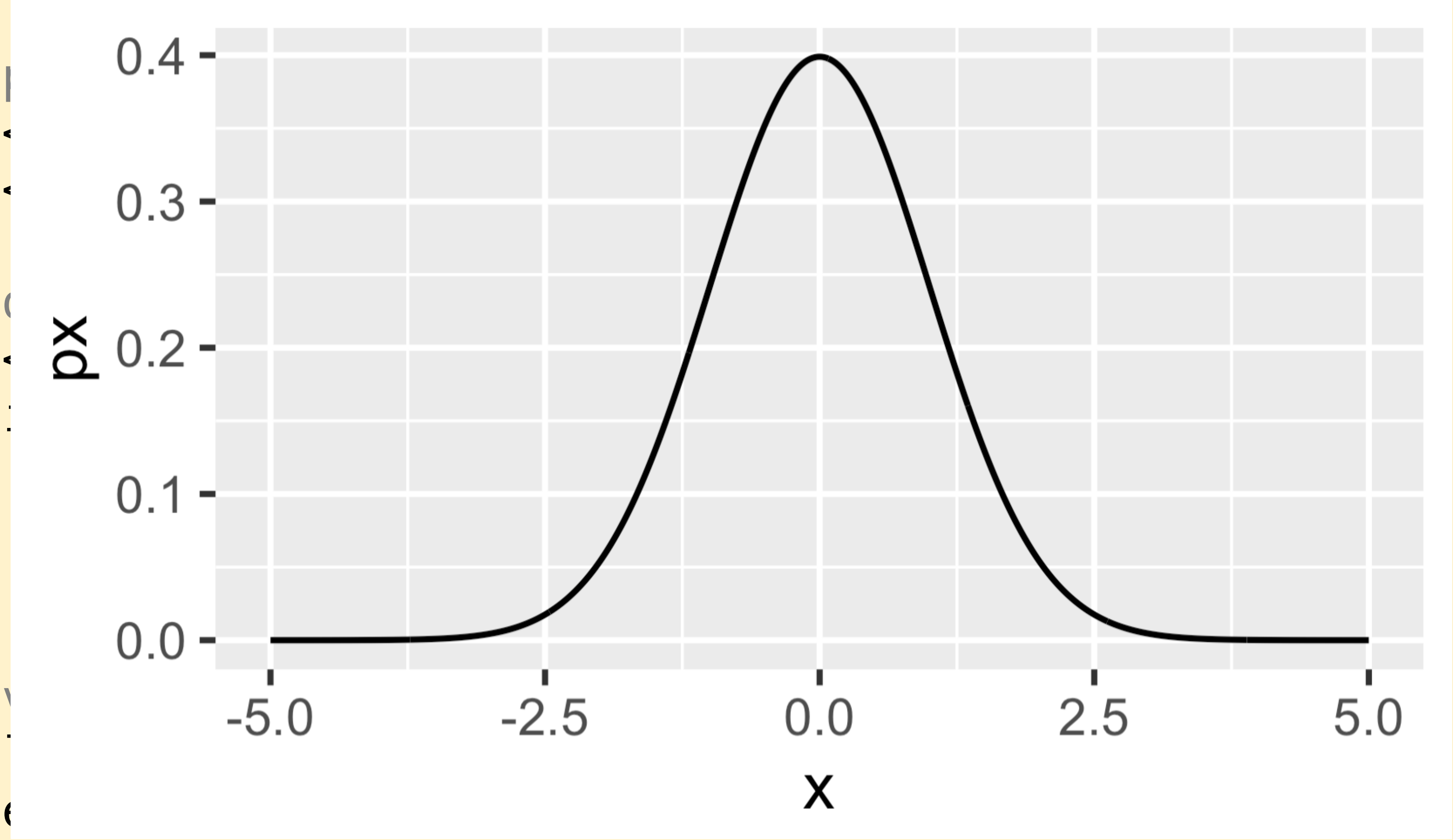
```
## packages ----
library(tidyverse)

## parameteres ----
mu <- 0
sd <- 1

## function ----
Norm_p <- function(x, mu, sd){
  exp(-(x - mu)^2 / 2 * sd^2) / sqrt(2 * pi * sd^2)
}

## data ----
df <-
  tibble(
    x = seq(-5, 5, by = 0.01),
    px = Norm_p(x = x, mu = mu, sd = sd)
  )
```

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

```r
## packages ----
library(tidyverse)

## parameteres ----
mu <- 0
sd <- 1

## data ----
df <-
  tibble(
    x = seq(-5, 5, by = 0.01),
    px = dnorm(x = x, mean = mu, sd = sd)
  )

## visualization ----
ggplot(data = df) +
  aes(x = x, y = px) +
  geom_path()
```

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

```
## packages ----
library(tidyverse)

## p
mu <
sd <

## d
df <
  ti


  )

## v
ggp
  ae
  geom_path()
```



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

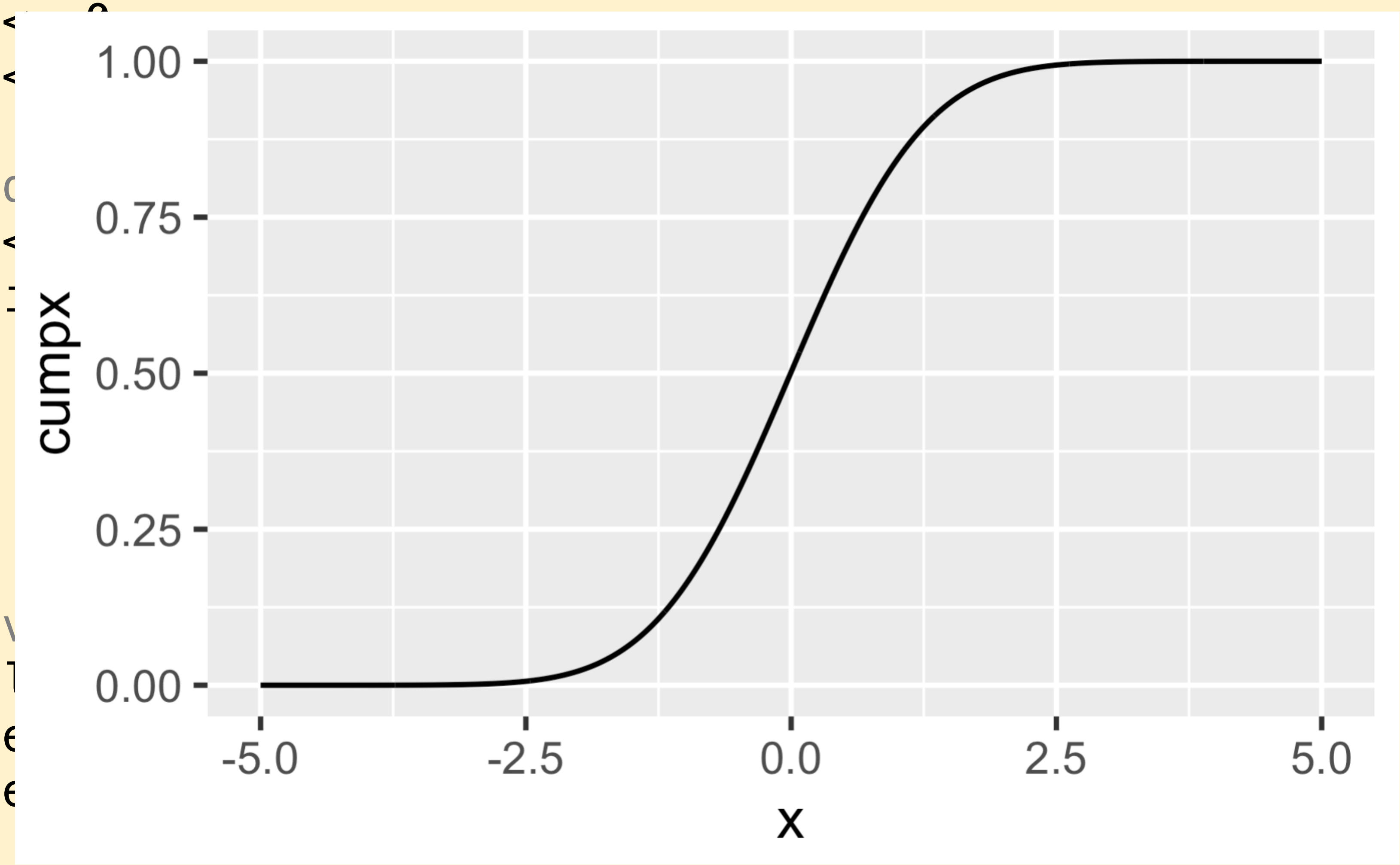
```r
## parameteres ----
mu <- 0
sd <- 1

## data ----
df <-
  tibble(
    x = seq(-5, 5, by = 0.01),
    px = dnorm(x = x, mean = mu, sd = sd), # 確率密度
    cumpx = cumsum(px) # 累積和
  )

## visualization ----
ggplot(data = df) +
  aes(x = x, y = cumpx) +
  geom_path()
```

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

```r
## parameteres ----
mu <- 0
sd <- 1

## data ----
df <-
  tibble(
    x = seq(-5, 5, by = 0.01),
    px = dnorm(x = x, mean = mu, sd = sd), # 確率密度
    cumpx = cumsum(px) * 0.01 # 積分値 = 累積和 * 幅
  )

## visualization ----
ggplot(data = df) +
  aes(x = x, y = cumpx) +
  geom_path()
```

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

```
## parameteres ----
mu <- 0
sd <

## c
df <
  ti

  )

## v
ggpl
  ae
  ge
```

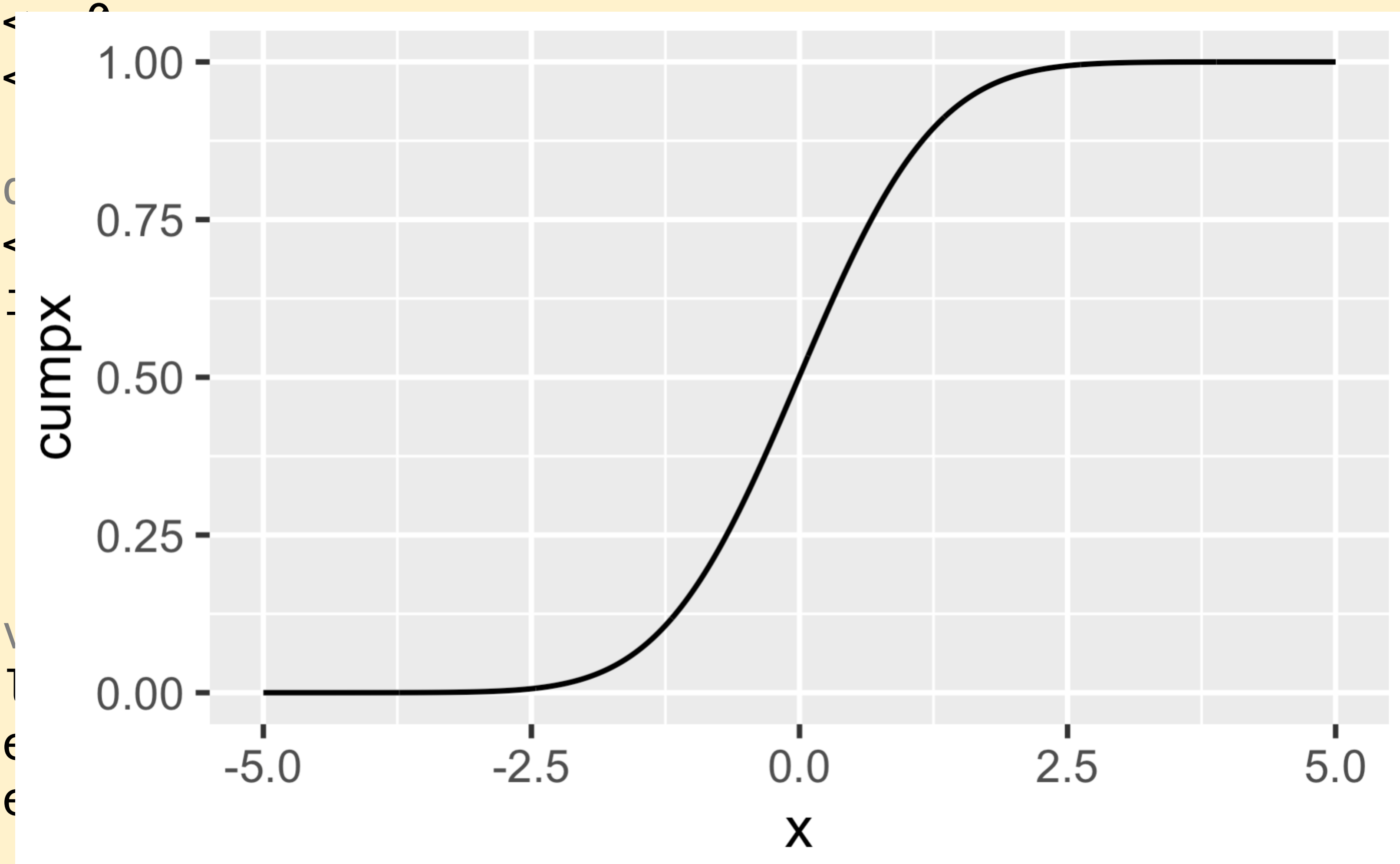


$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

```r
## parameteres ----
mu <- 0
sd <- 1

## data ----
df <-
  tibble(
    x = seq(-5, 5, by = 0.01),
    px = dnorm(x = x, mean = mu, sd = sd), # 確率密度
    cumpx = pnorm(q = x, mean = mu, sd = sd) # 累積確率
  )

## visualization ----
ggplot(data = df) +
  aes(x = x, y = cumpx) +
  geom_path()
```

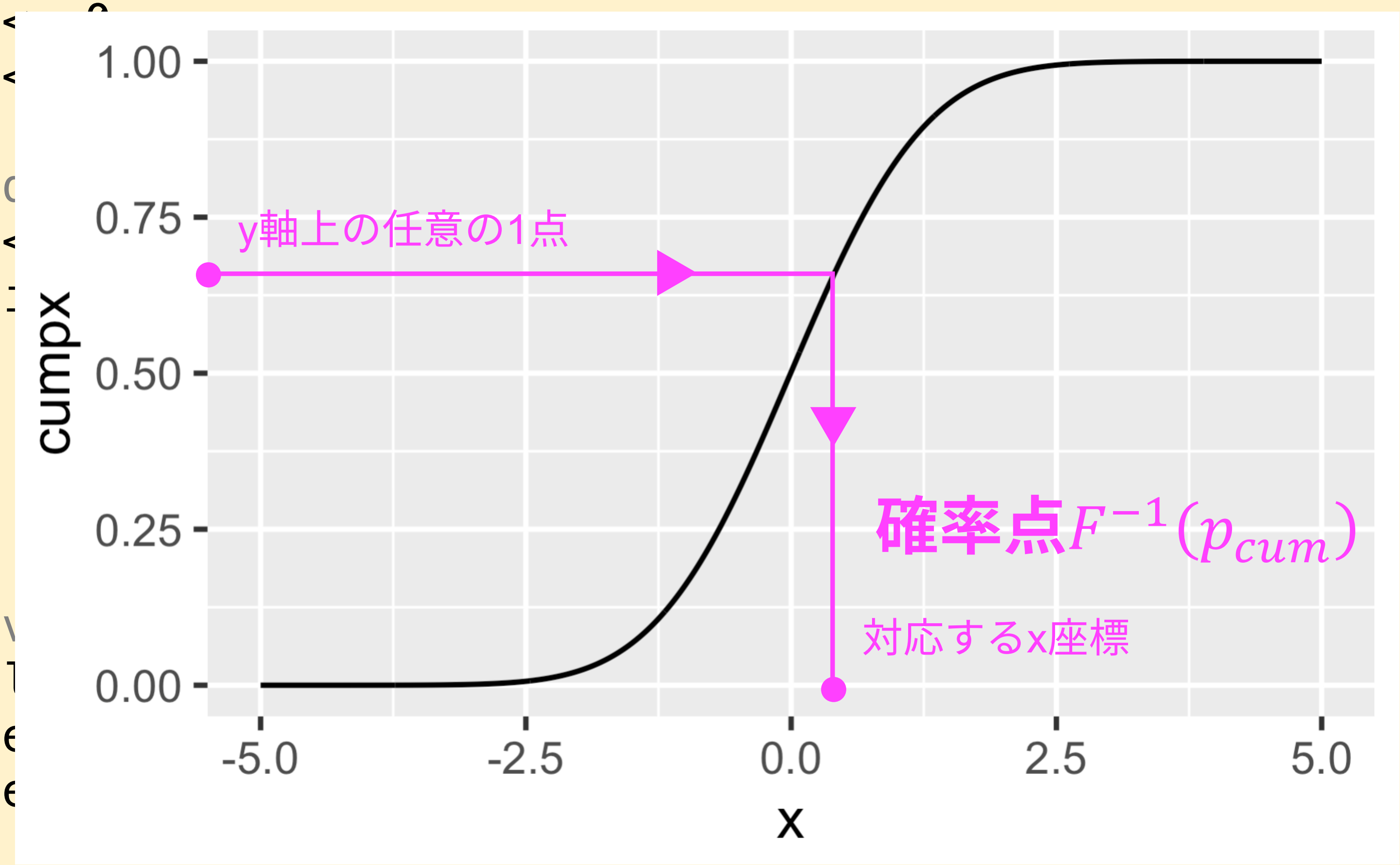
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

```
## parameteres ----
mu <
sd <

## c
df <
  ti



  )

## v
ggpl
  ae
  ge
```



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

```
## parameteres ----
mu <
sd <

## c
df <
  t:


  )

## v
ggpl
  ae
  ge
```

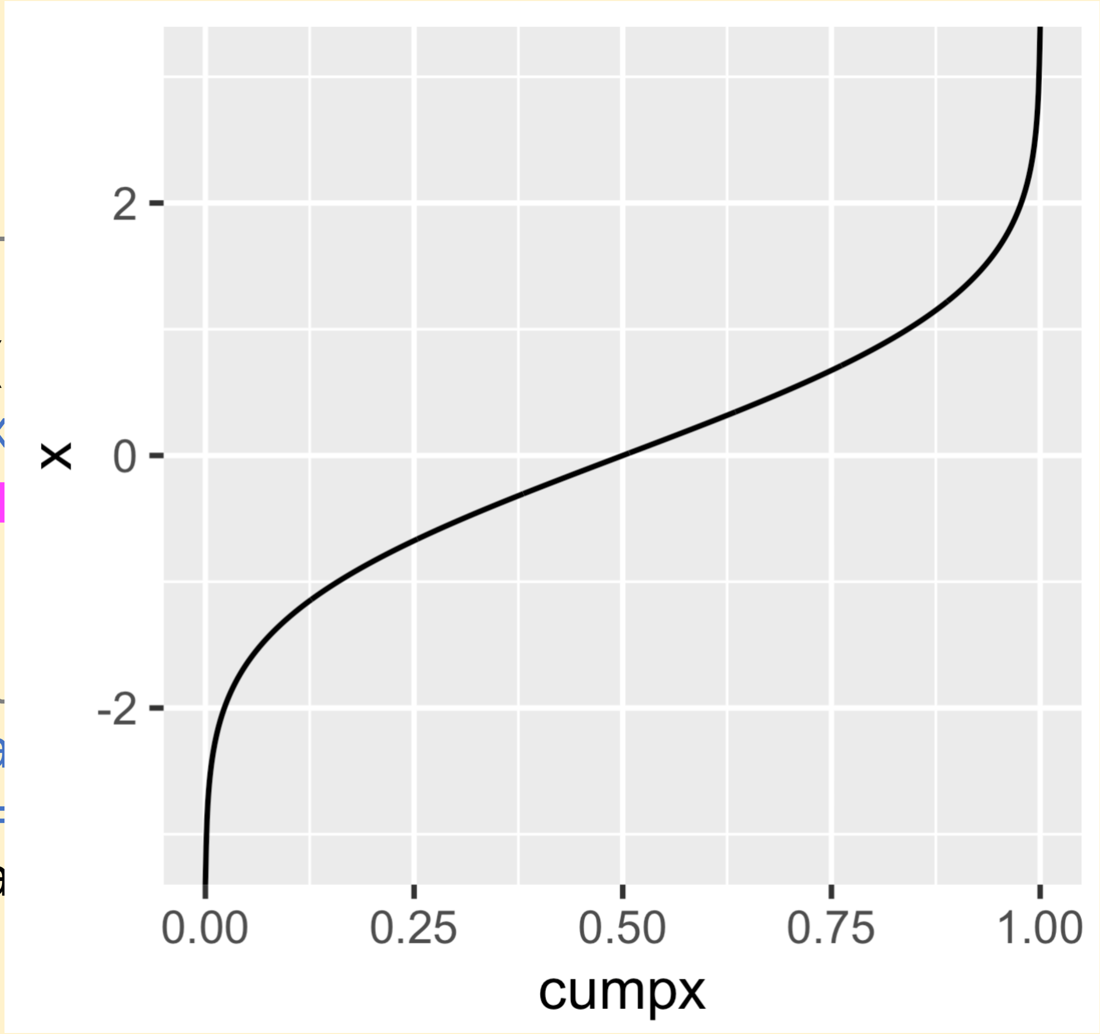


$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

```r
## parameteres ----
mu <- 0
sd <- 1

## data ----
df_qnorm <-
  tibble(
    cumpx = seq(0, 1, by = 0.001),
    x = qnorm(p = cumpx, mean = mu, sd = sd), # 確率点
  )

## visualization ----
ggplot(data = df_qnorm) +
  aes(x = cumpx, y = x) +
  geom_path()
```

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

```
## parameteres ----
mu <- 0
sd <- 1

## data -
df_qnorm
  tibble(
    cumpx
    x = q                                    ), # 確率点
  )

## visual
ggplot(da
  aes(x =
  geom_pa
```



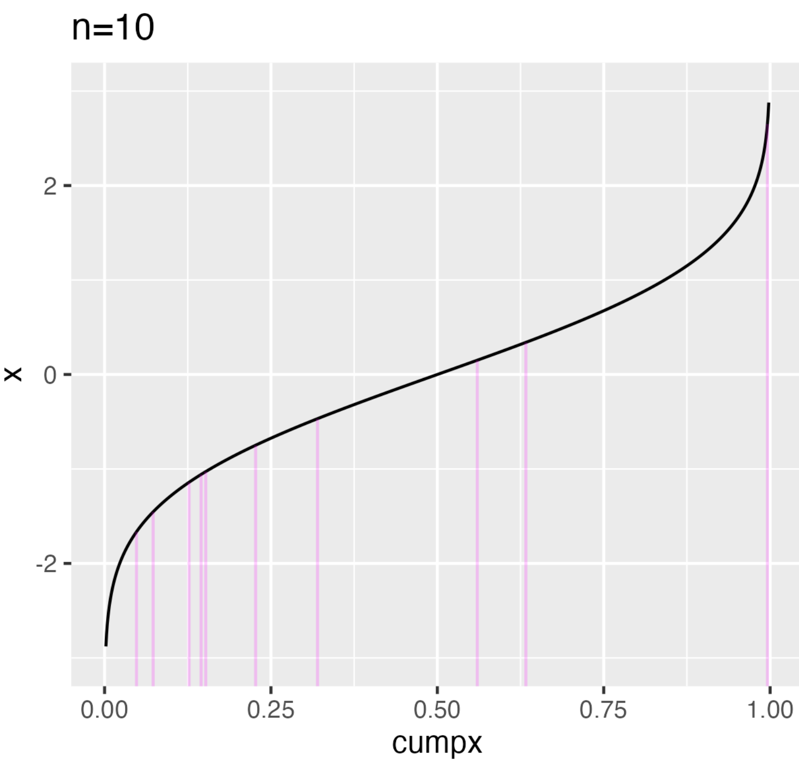
確率点のデータ ⟶ 10行を無作為抽出

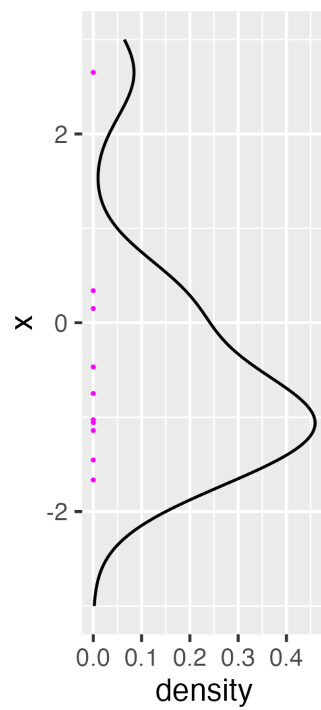
```
df_qnorm <-
  tibble(
    cumpx = seq(0, 1,
                by = 0,001),
    x = qnorm(p = cumpx,
              mean = mu,
              sd = sd))
```

```
df_sample_qnorm <-
  df_qnorm %>%
  sample_n(size = 10,
           replace = TRUE)
```

重複抽出を許可
(復元抽出)

密度分布をプロット
geom_density()関数

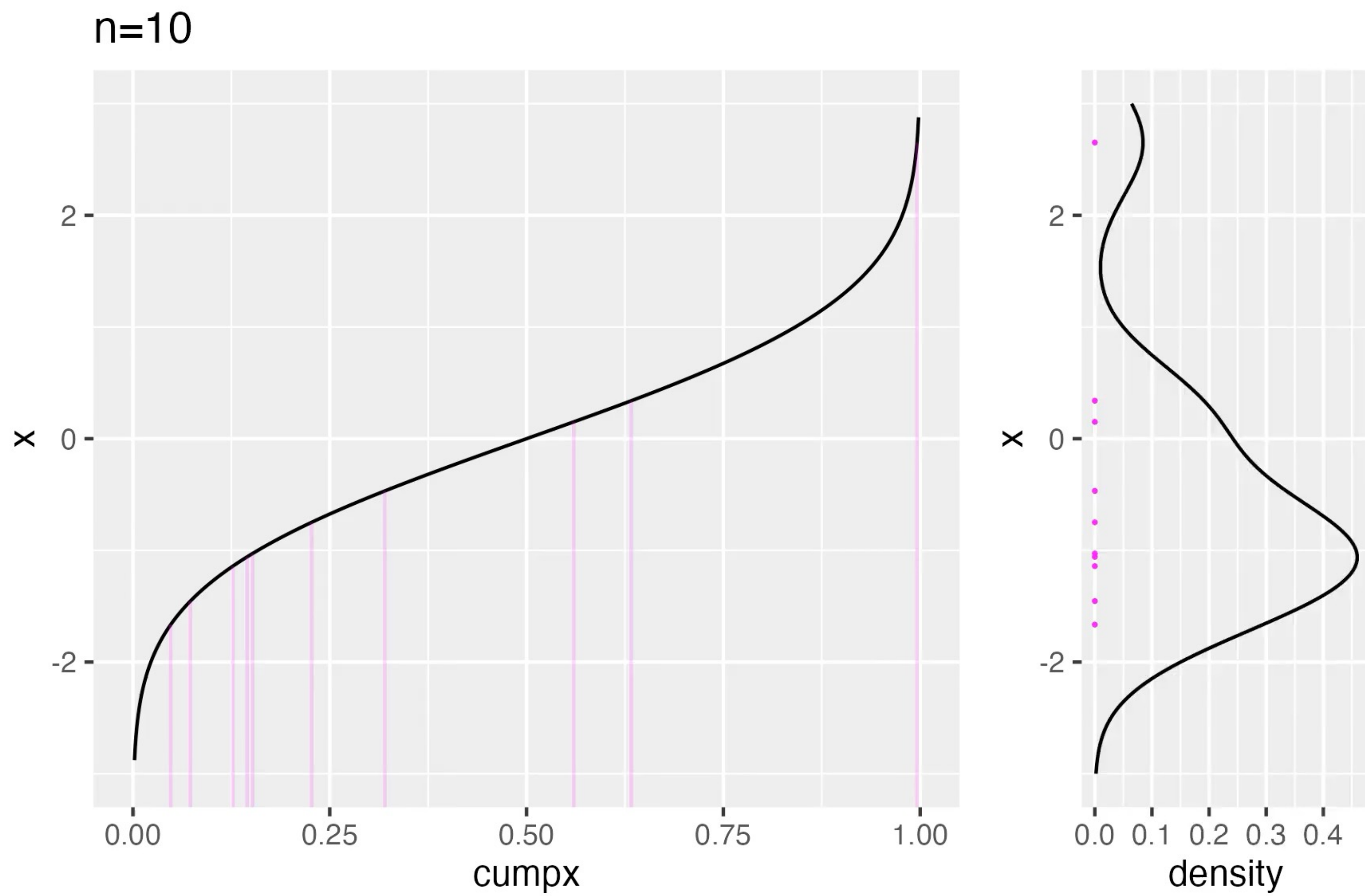




確率点

$$x = F^{-1}(p_{cum})$$

無作為抽出された
xの密度分布

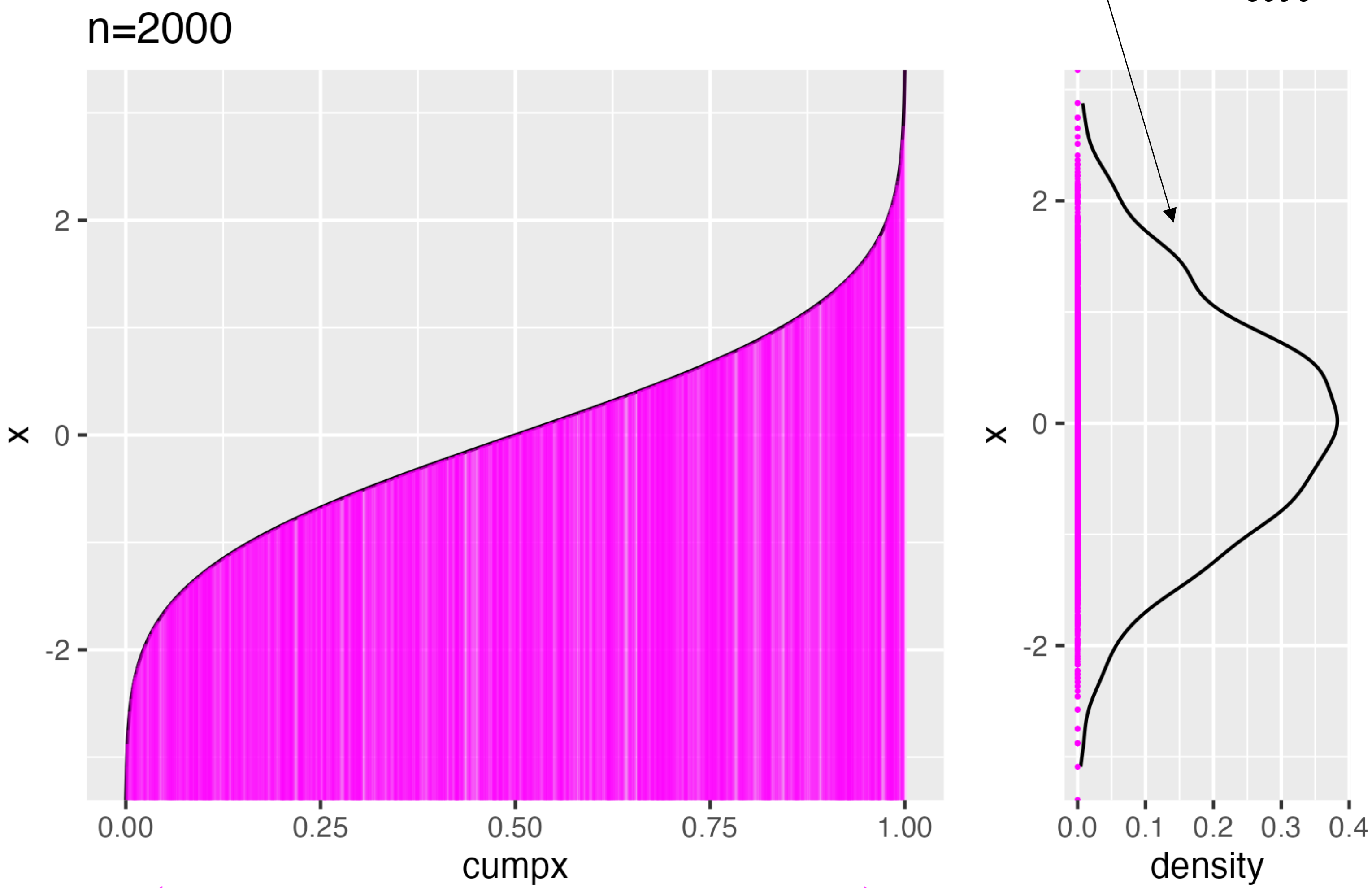


確率点(累積確率の逆関数)

$$x = F^{-1}(p_{cum})$$

確率密度

$$\mathrm{f}(x) = \frac{d}{dx}F(x)$$



$$Unif(\min = 0, \max = 1)$$

確率点（累積確率の逆関数）

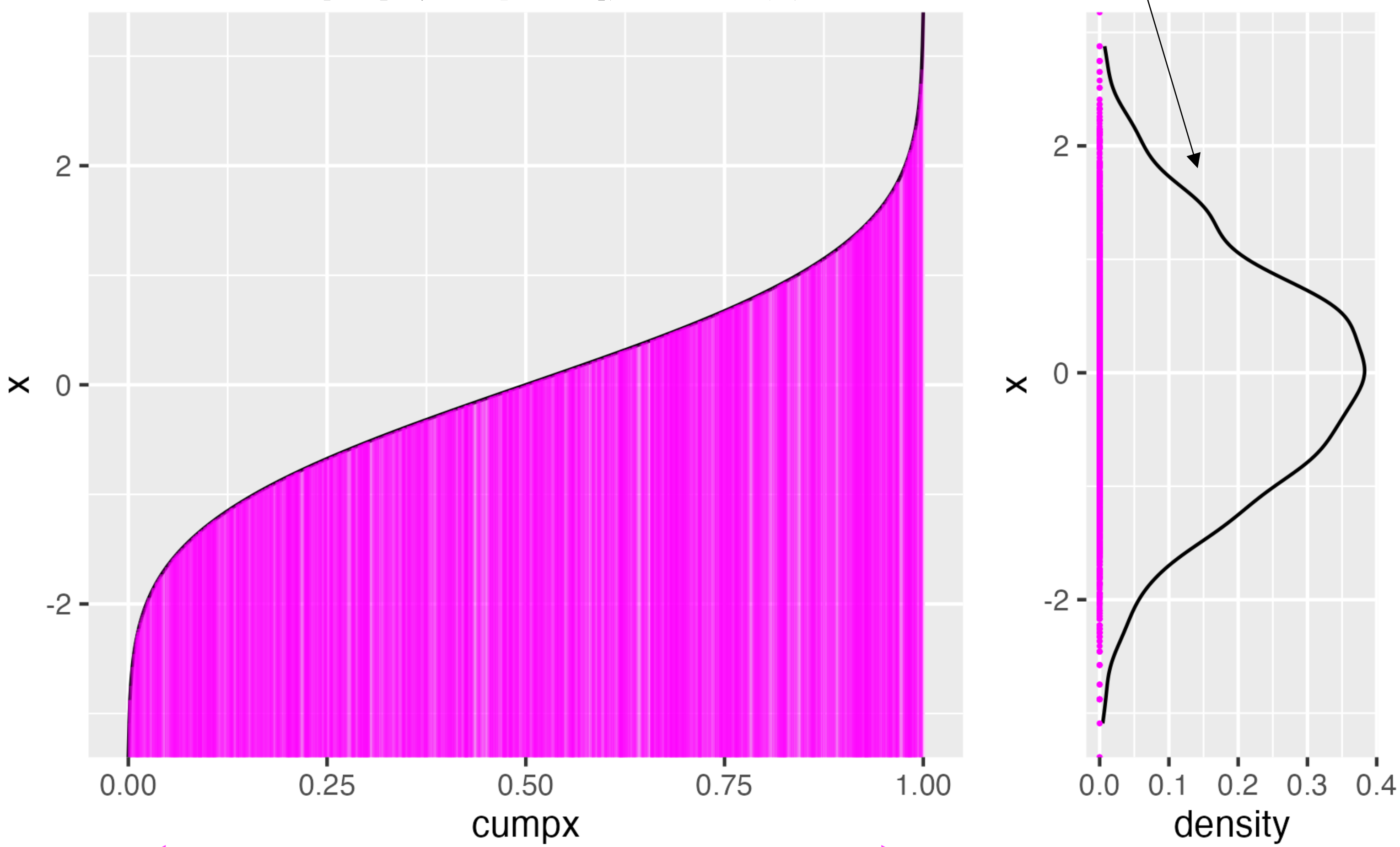
$$x = F^{-1}(p_{cum})$$

確率分布に従う乱数

確率密度

$$\mathrm{f}(x) = \frac{d}{dx}F(x)$$

n=2000



$$Unif(\mathrm{min} = 0, \mathrm{max} = 1)$$

確率点(累積確率の逆関数)

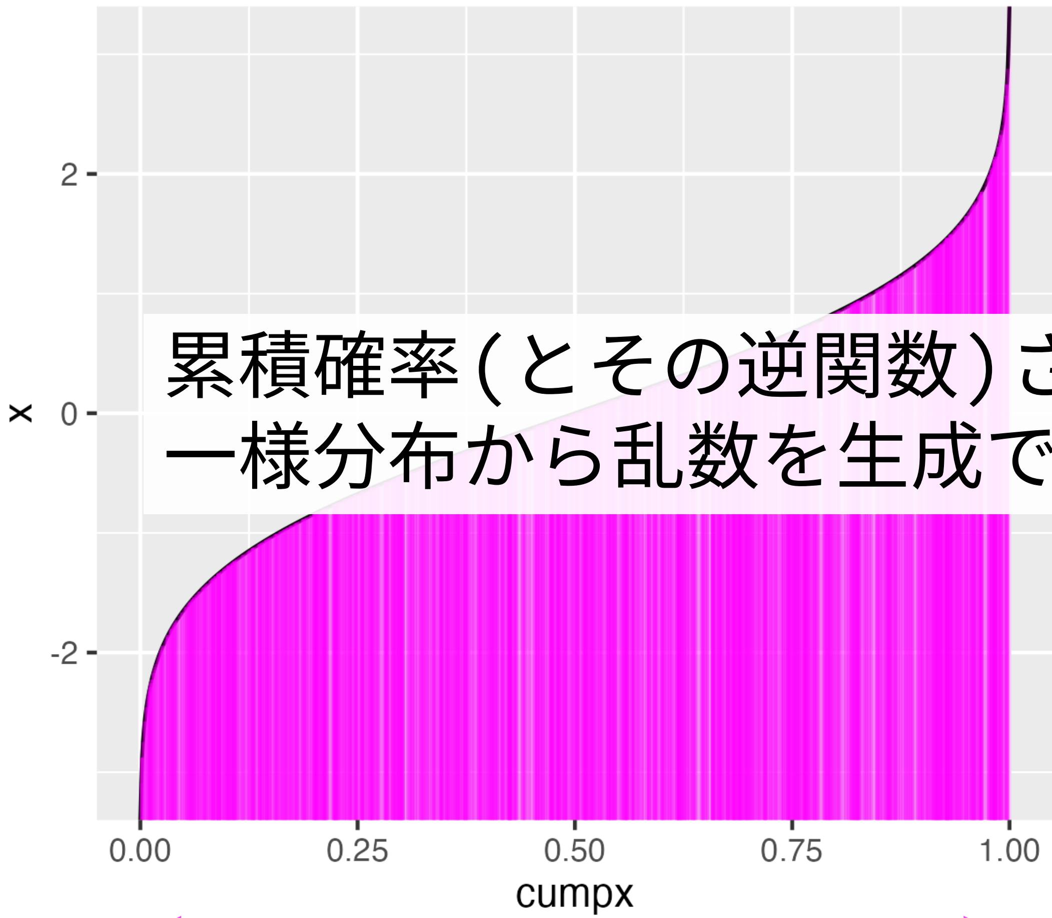
$$x = F^{-1}(p_{cum})$$

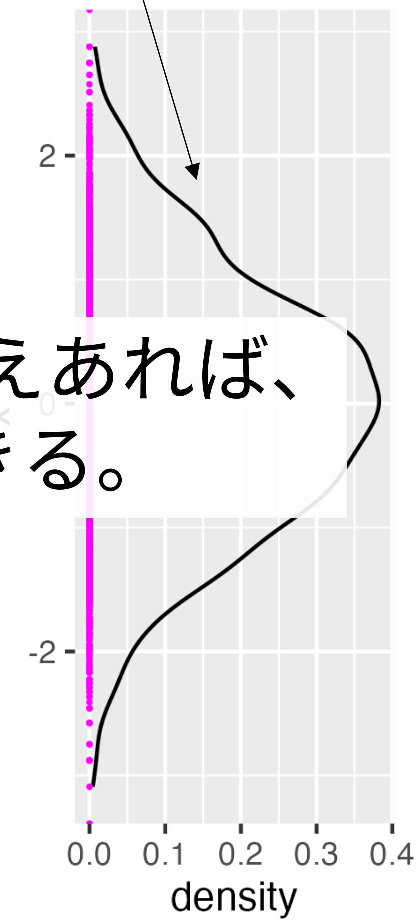
確率分布に従う乱数

確率密度

$$f(x) = \frac{d}{dx}F(x)$$

n=2000

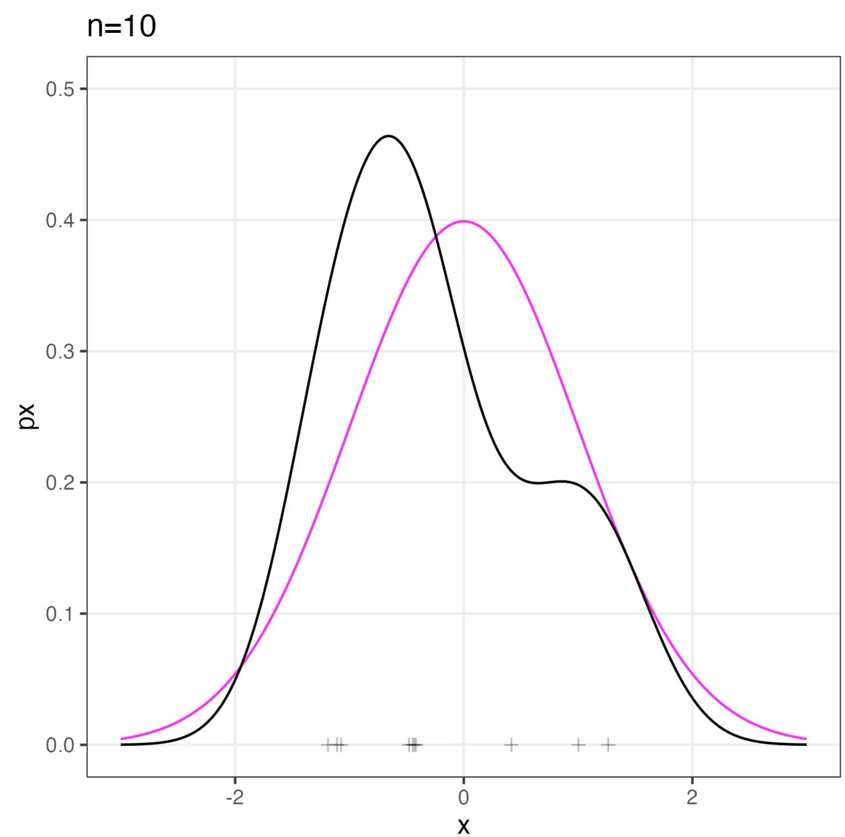


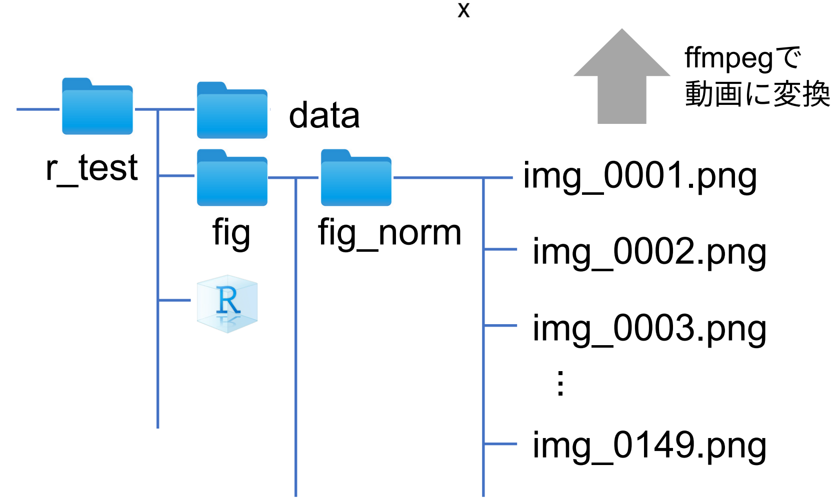


累積確率(とその逆関数)さえあれば、
一様分布から乱数を生成できる。

$$Unif(\mathrm{min} = 0, \mathrm{max} = 1)$$

```
## env
library(tidyverse)
set.seed(71)
dir.create("fig/fig_norm")

## parameters
n <- 3000
mu <- 0
sd <- 1

## data ----
df_rnorm <-
  tibble(tag = seq(1, n, by = 1),
         x = rnorm(n = n, mean = mu, sd = sd))

df_dnorm <-
  tibble(x = seq(-3, 3, by = 0.001),
         px = dnorm(x = x, mean = mu, sd = sd))

## visualization ----
k <- c(seq(10, 500, by = 5), seq(550, 3000, by = 50))

for(i in 1:length(k)){
  df_rnorm_i <-
    df_rnorm %>%
    filter(tag <= k[i])

  g <-
    ggplot(data = df_rnorm_i) +
    aes(x = x) +
    geom_path(data = df_dnorm,
              aes(y = px),
              color = "magenta") +
    geom_density() +
    geom_point(y = 0, shape = "plus", alpha = 0.3) +
    coord_cartesian(xlim = c(-3, 3)) +
    scale_y_continuous(limits = c(0, 0.5)) +
    labs(title = str_c("n=", k[i])) +
    theme_bw() +
    theme(panel.grid.minor = element_blank())

  ## save
  i %>%
    formatC(width = 4, flag = "0") %>%
    str_c("fig/fig_norm/img_", ., ".png") %>%
    ggsave(g, width = 5, height = 5)

  cat(i, " ")
}
```





set.seed()関数の話

```r
runif(n = 1, min = 0, max = 1)
```

```r
set.seed(42)
runif(n = 1, min = 0, max = 1)
```

```r
set.seed(71)
runif(n = 1, min = 0, max = 1)
```

- 確率の話
  - 確率論の基礎
  - 確率分布 in R
- 回帰モデルの話