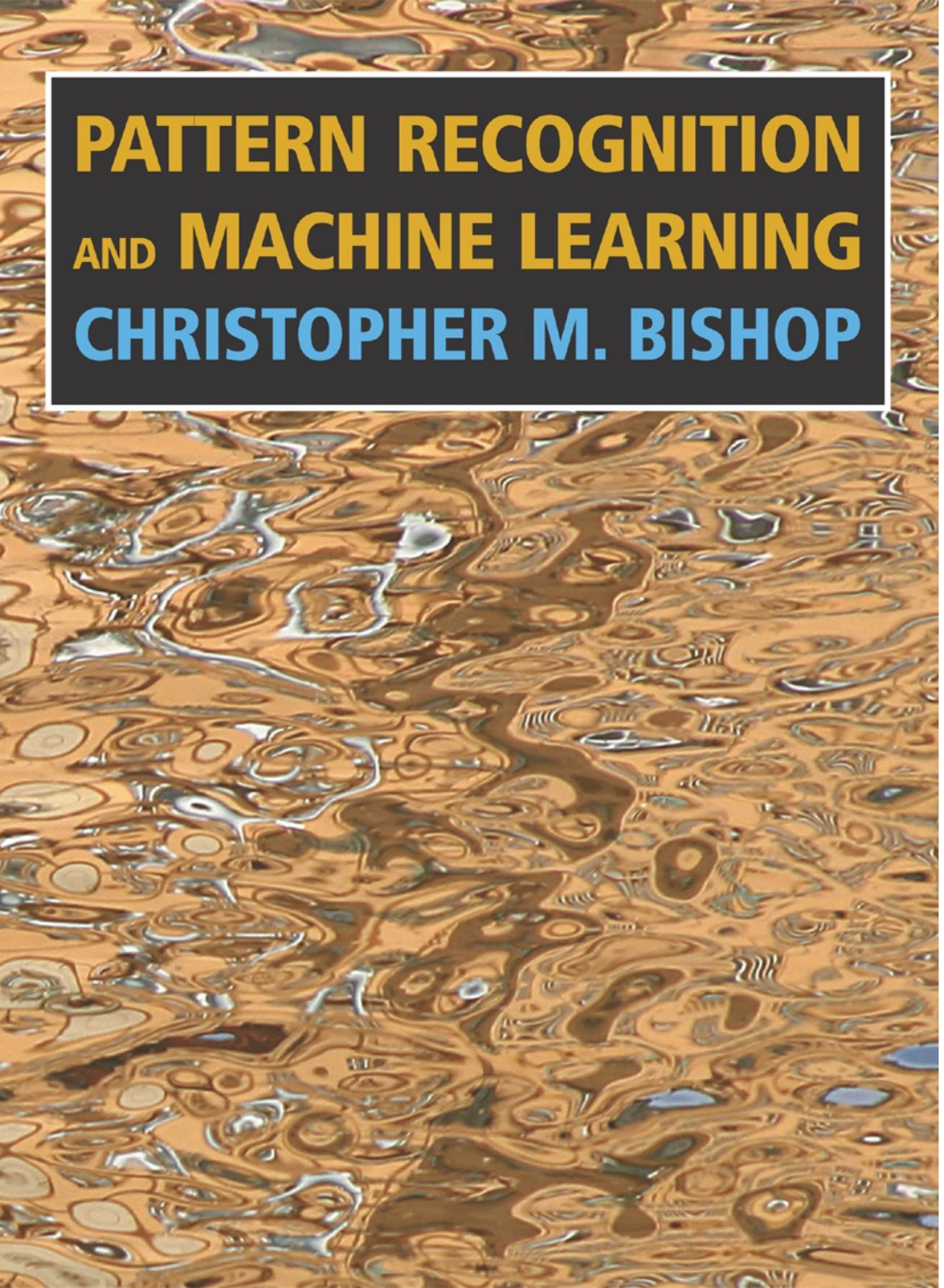
Introduction

正田 備也

masada@rikkyo.ac.jp



参考書

C. Bishop. *Pattern Recognition and Machine Learning*. 2006.

- この授業はベイズ的な統計モデリングの授業です
- 他にもいろいろな文献を参考にしつつ内容を構成しています

参考になる講義資料

- 渡辺澄夫先生の講義資料
  - http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/da2020.html
  - http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/joho-gakushu6.html

    - 私の講義の内容は、上記の講義資料の内容よりも、初歩的です。

今日のお題

- 統計モデリングとは
- 統計モデリングを機械学習の世界の中に位置づける
  - unsupervised / supervised
  - generative / discriminative
- 確率の復習
  - 確率変数、同時確率、周辺化、条件付き確率、ベイズ則、確率分布、期待値・・・

統計モデリングとは

参考資料

http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/bayes000.pdf

統計的推測とは

- 私が関心のあるデータは、何らかの分布から生成されている。
  - これを「真の分布」と呼ぶ。
- その分布から生成されたとみなせるデータ集合が手元にある。
- そこで、ある定められた手続きにしたがって、
- その手元にあるデータから、真の分布を推測する。
- これを、統計的推測と呼ぶ。

この授業で説明する統計的推測の方法

- 最尤法 (MLE; maximum likelihood estimation)
  - 最尤推定
- MAP法 (MAP; maximum a posteriori probability estimation)
  - 事後確率最大化推定
- ベイズ法 (Bayesian inference)
  - ベイズ推論

統計学は不良設定問題を扱う学問

- 前のスライドに示したどの方法を使っても・・・

推測≠真の分布

  - つまり、推測は常に間違う。

- では、真の分布を推測しても無意味なのか？
  - 推測がいつも間違うとしても・・・
  - どのくらい間違っているかを、統計学で明らかにできる！

統計モデリングとは

http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/bayes010.pdf

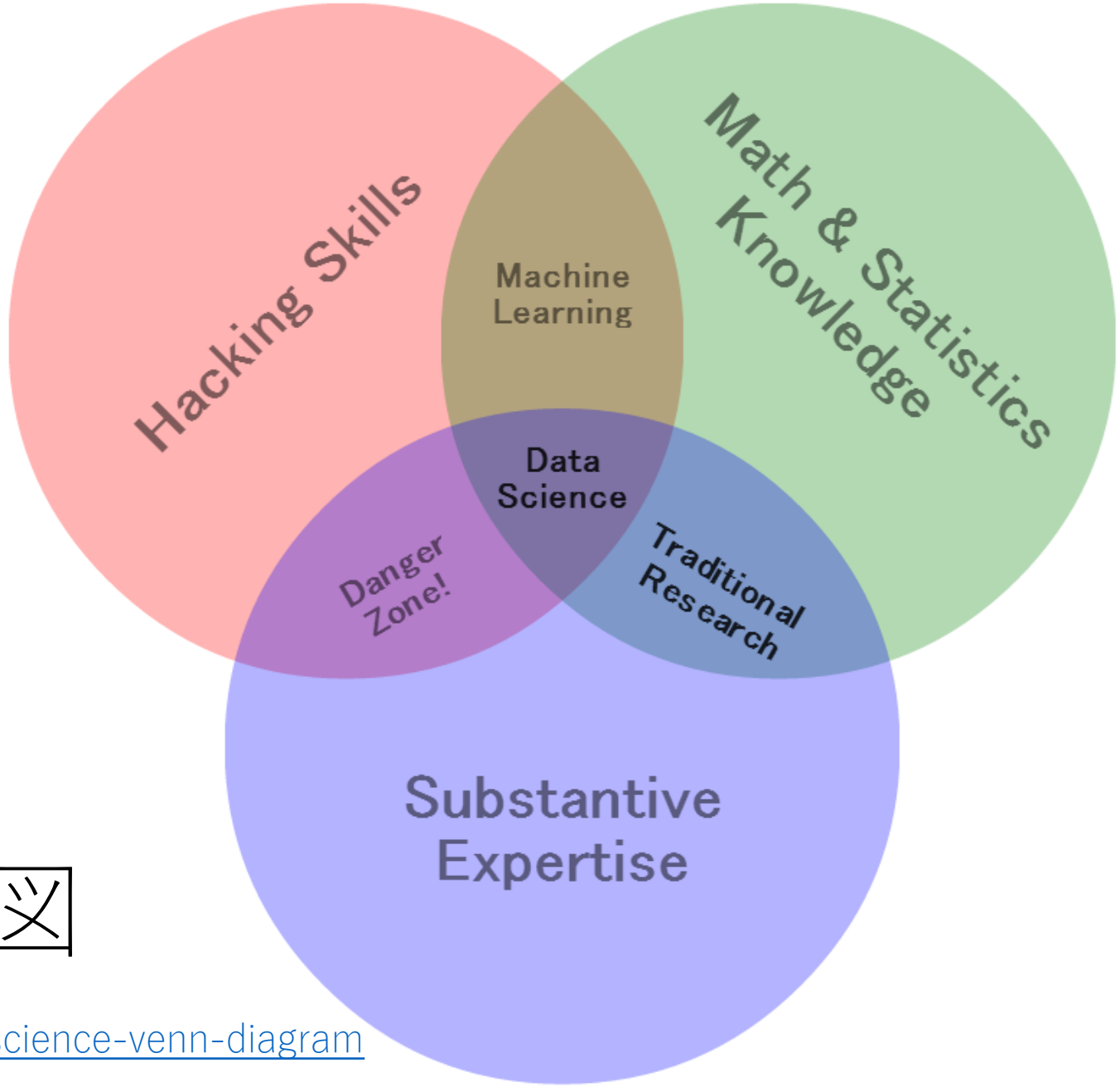
- データを生成する真の分布は、未知。
- そこで、我々は、データを生成する分布を、自由に設定する。
- そして、推測された分布と、真の分布とが、どのくらい違っているかを、数理的に明らかにする。
  - 「どのくらい違っているか」＝汎化誤差
  - 汎化誤差を推測できる方法を作ることが統計モデリングの目標

ベイズ的統計モデリングの実験の手続き

- 真の分布から生成されたとみなせるデータ集合を入手する。
- モデルを設定する。
  - データをモデリングする分布と、事前分布を決める。
- 事後分布を（近似的に）求める。
- 予測分布を（近似的に）求める。
- テストデータ上で汎化誤差を（近似的に）求める。
  - 異なるモデリング間で汎化誤差を比べれば、どちらが良いか分かる。

統計モデリングを
機械学習の世界の中に位置づける

データ科学のベン図



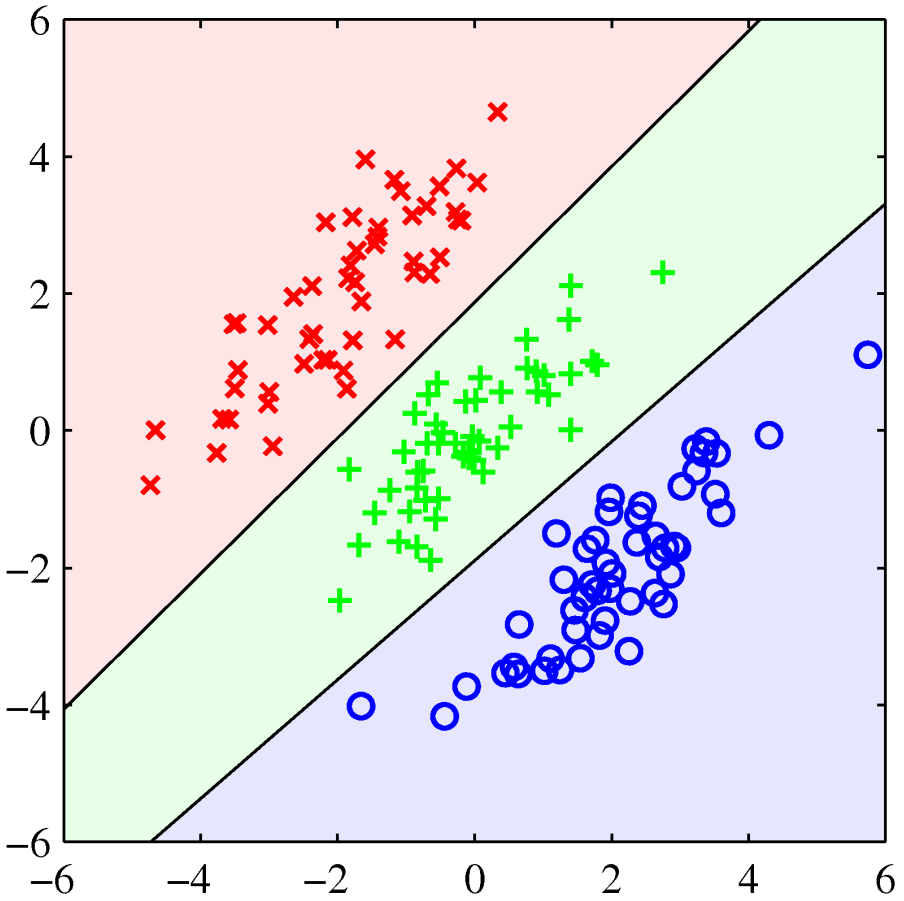
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

機械学習の2区分（既習）

- 教師あり学習 (supervised learning)
  - classification
  - regression
- 教師なし学習 (unsupervised learning)
  - clustering
  - dimensionality reduction

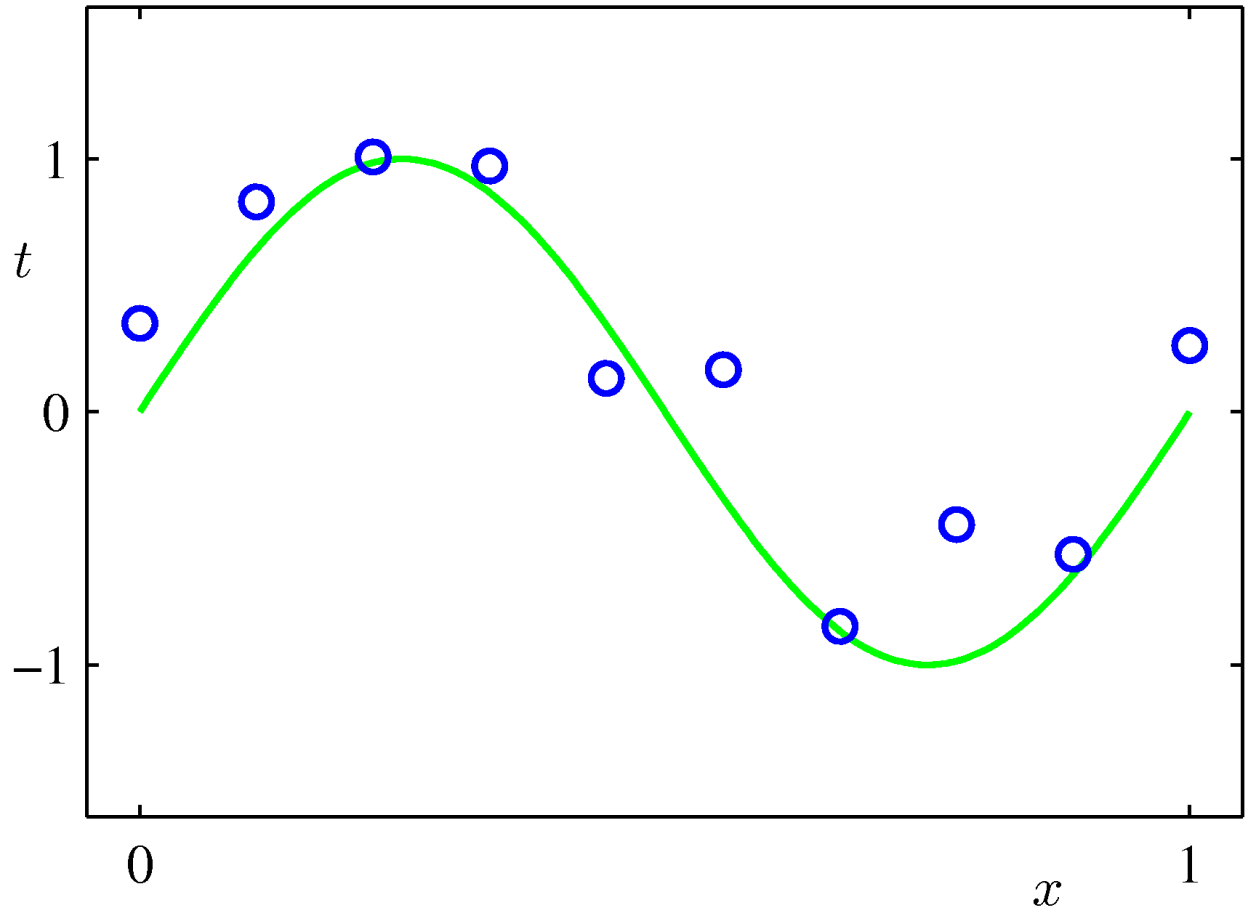
強化学習は
これらとは
別枠です。

classificationのイメージ



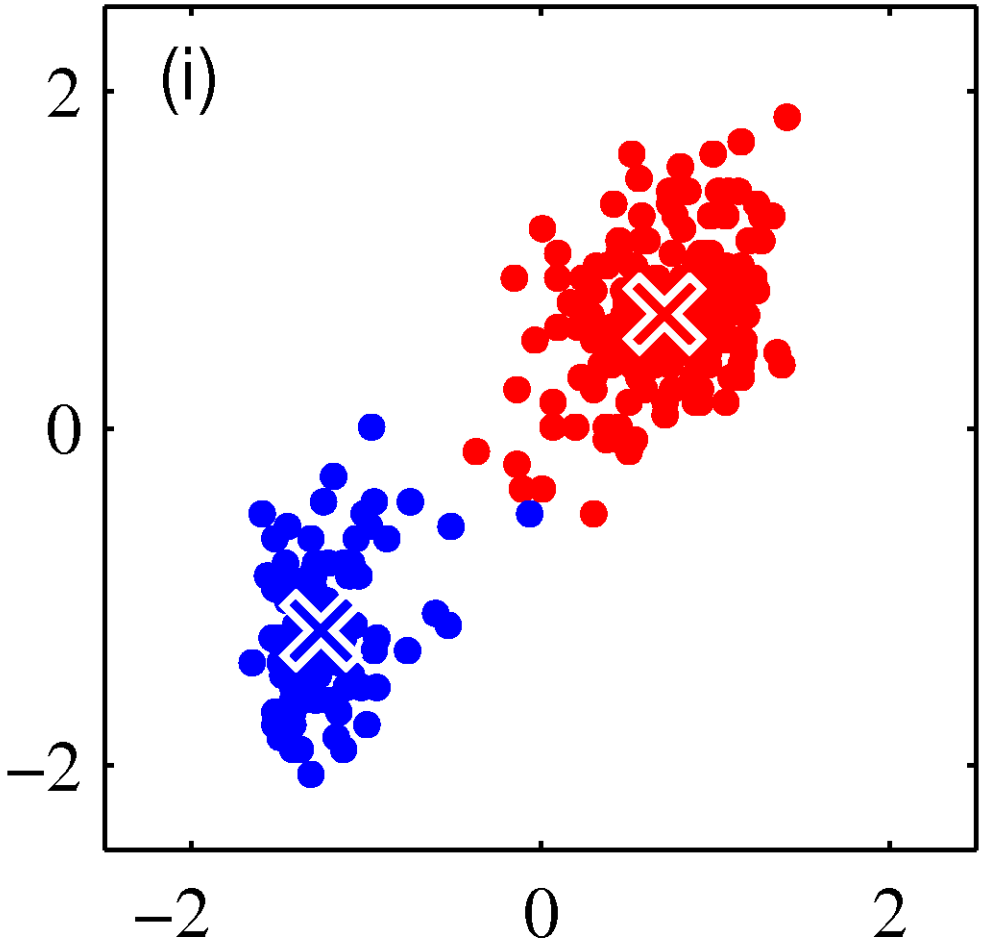
C. Bishop. *Pattern Recognition and Machine Learning*. 2006.

regressionのイメージ



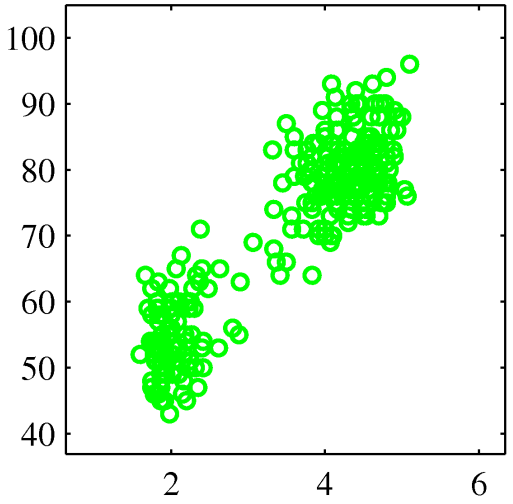
C. Bishop. *Pattern Recognition and Machine Learning*. 2006.

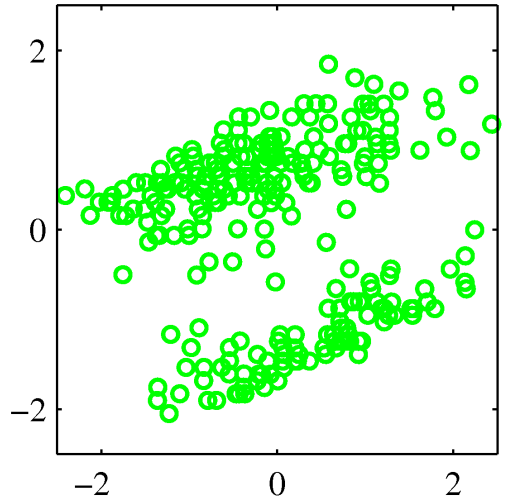
clusteringのイメージ

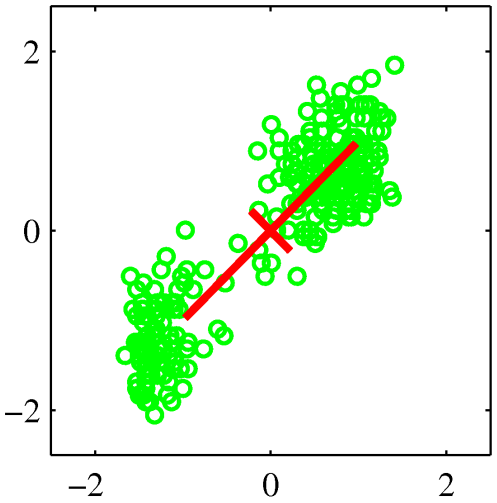


C. Bishop. *Pattern Recognition and Machine Learning*. 2006.

dimensionality reductionのイメージ







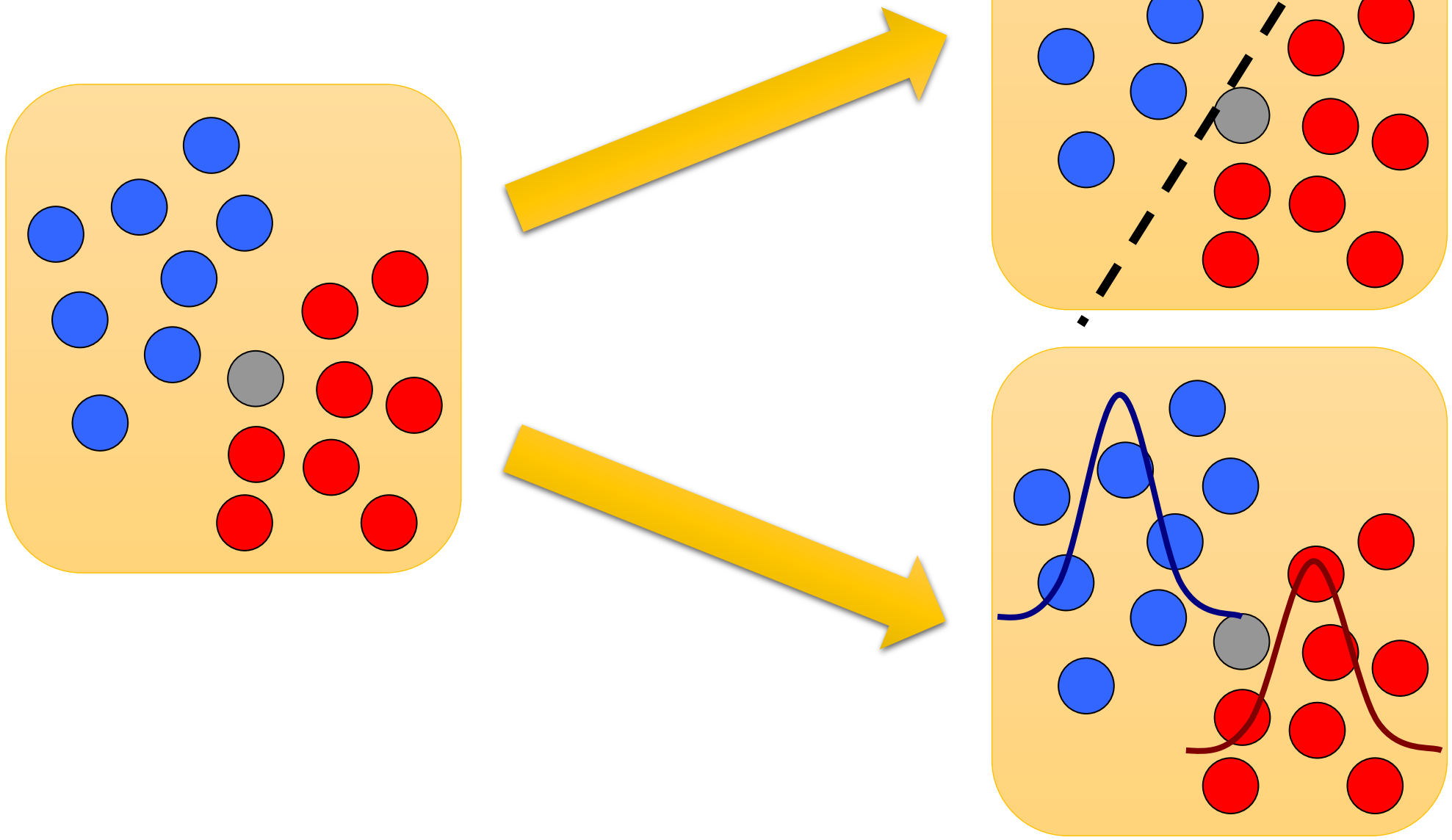
C. Bishop. *Pattern Recognition and Machine Learning*. 2006.

機械学習の別の2区分

http://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf

- クラスを識別する境界を見つける
  - 境界から遠いところにあるデータは気にしない
- クラスを生成する分布を見つける
  - 境界はこの分布 distribution の違いから派生する

boundary vs distribution



discriminative vs generative

- discriminative methods
  - クラスを識別する境界を計算機に学習させる
- generative methods
  - クラスを生成する確率分布を計算機に学習させる
    - この授業ではこちらを教えます。

generative approaches (1/2)

‘Generative models in machine learning posit that there is some underlying (…) process that is generating the data you are observing and aim to use the data to infer the parameters of that underlying process, which then lets you classify the data.’

http://www.forbes.com/sites/quora/2015/02/12/what-is-the-future-of-machine-learning/

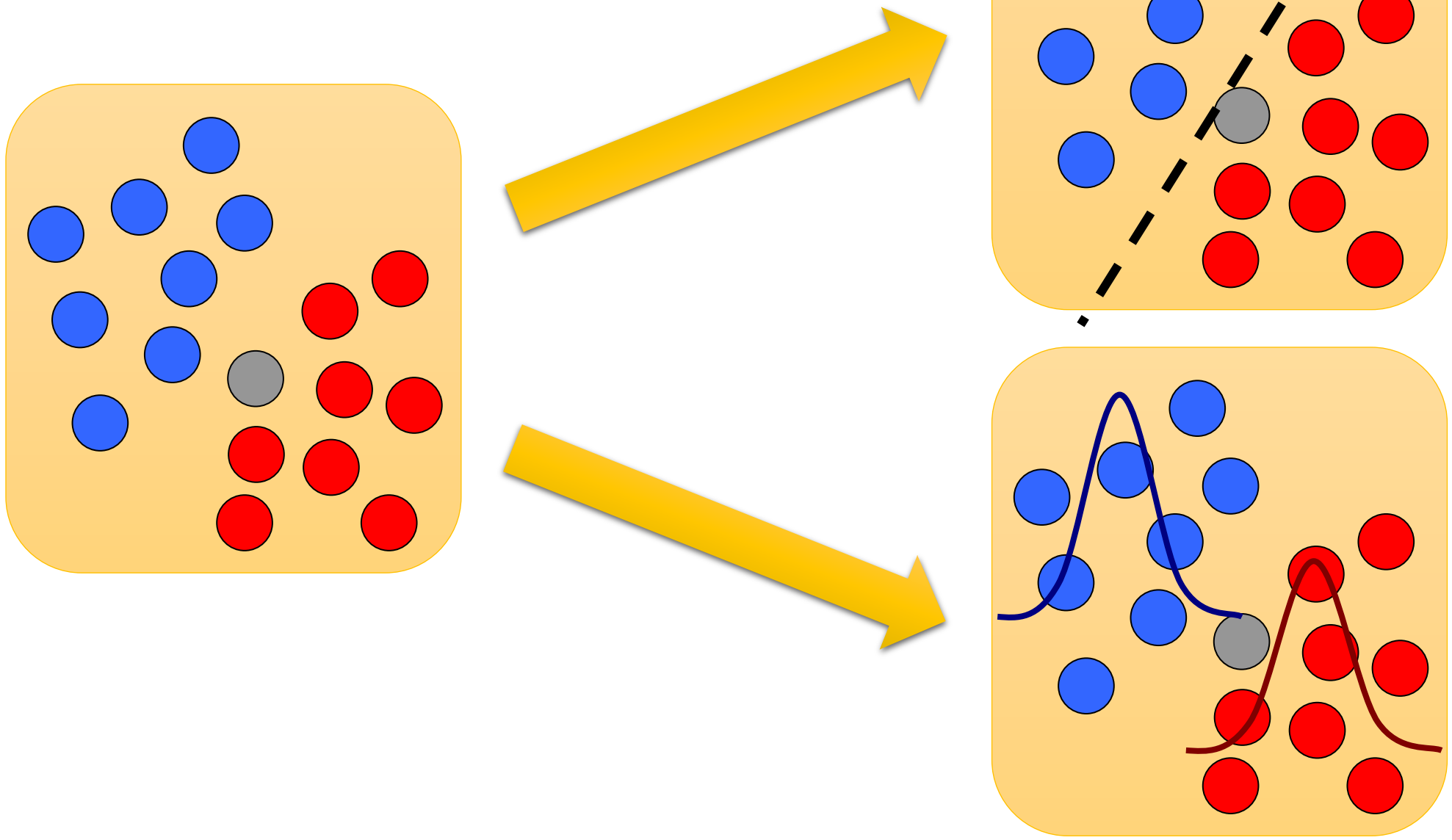
generative approaches (2/2)

‘In my mind, if you succeed in solving a generative model, you have “understood” the data and the problem.’

‘Unfortunately for me, discriminative models tend to work better than generative models to solve lots of machine learning problems’

http://www.forbes.com/sites/quora/2015/02/12/what-is-the-future-of-machine-learning/

boundary vs distribution



データを生成する確率分布を推定することの難しさ

http://ibisml.org/archive/ibisml001/Sugiyama.pdf

「統計的機械学習のほとんどの課題は、データの生成確率分布の推定を介して解決することができる。

しかし、確率分布の推定は機械学習における最も困難な問題の一つとして知られているため、現実的には分布推定を回避しながら対象となる課題を解決することが望ましい。」(2010年)

GAN [Goodfellow+ 14]

- 観測データを生成する確率分布を推定する手法

1. 観測データの分布とモデルの出力の分布の密度比を推定する
   - 正規化されていないので分布の形しか分からないため
   - データがどちらから来たかを予測する識別モデルを学習
     - [Goodfellow+ 14]のProposition 1
2. 推定された密度比を1に近づける
     - [Goodfellow+ 14]のTheorem 1

Diffusion models [Sohl-Dickstein+ 2015]

- Lilian Wengさんのブログ記事
  - https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

確率の復習

確率の復習

- 確率変数 random variable
- 同時確率 joint probability
- 周辺化 marginalization
- 条件付き確率 conditional probability
- ベイズ則 Bayes rule
- 確率分布 probability distribution
- 期待値 expectation

確率変数 random variable

- 確率変数: どの値をとるかが不確かuncertainな変数
- $p(x = v)$ は、確率変数xが値vをとる確率、という意味
- $\sum_x p(x) = 1$を満たす
  - つまり、xがとりうる値を$v_1, v_2, \ldots, v_W$とすると、以下を満たす

$$p(x = v_1) + p(x = v_2) + \cdots + p(x = v_W) = 1$$

    - このように、「$\sum_x p(x) = 1$」という書き方は、省略を含んでいる
      - こういうのにも慣れましょう

同時確率 joint probability

$$p(x = v, z = c)$$

- 上は、確率変数xが値vをとり、かつ、確率変数zが値cをとる確率
- $\sum_x \sum_z p(x, z) = 1$を満たす。つまり・・・
- 確率変数zがとりうる値を$c_1, c_2, \ldots, c_K$とすると、以下が成り立つ

$$p(x = v_1, z = c_1) + \cdots + p(x = v_1, z = c_K)$$
$$+p(x = v_2, z = c_1) + \cdots + p(x = v_2, z = c_K)$$
$$+ \cdots + p(x = v_W, z = c_1) + \cdots + p(x = v_W, z = c_K) = 1$$

周辺化 marginalization

$$p(x) = \sum_z p(x, z)$$

- $p(x)$は$p(x, z)$から上の計算で得ることができる。つまり・・・

$$p(x = v) = p(x = v, z = c_1) + \cdots + p(x = v, z = c_K)$$

- 特定の変数がとりうるすべての値についてその確率を足し合わせることを、その変数を周辺化marginalizationすると言う
- 得られた確率$p(x)$を、周辺確率と言う

例）ハンバーガー

	体重が 標準体重以下	**体重が 標準体重以上**
ハンバーガーを 毎日食べる	2	8
ハンバーガーを 毎日は食べない	38	2

例）確率変数を使って書くと…

	$x = a$	$x = b$
$z = s$	2	8
$z = t$	38	2

問題1-1

- $p(x = a)$を求めよ。
- $p(z = t)$を求めよ。
- $p(x = a, z = s)$を求めよ。
- $p(x = a, z = t)$を求めよ。

	$x = a$	$x = b$
$z = s$	2	8
$z = t$	38	2

例）それぞれの場合の確率を求めると…

	$x = a$	$x = b$
$z = s$	$\frac{1}{25}$	$\frac{4}{25}$
$z = t$	$\frac{19}{25}$	$\frac{1}{25}$

条件付き確率 conditional probability

$$p(x = v|z = c)$$

- 上は、確率変数zが値cをとるという事実が所与のとき、確率変数xが値vをとる確率（定義は下記のとおり）

$$p(x = v|z = c) \equiv \frac{p(x = v, z = c)}{p(z = c)}$$

- $\sum_x p(x|z = c) = 1$を満たす

問題1-2

- $p(x = a|z = s)$を求めよ。
- $p(x = a|z = t)$を求めよ。
- $p(z = s|x = a)$を求めよ。
- $p(z = t|x = a)$を求めよ。

	$x = a$	$x = b$
$z = s$	2	8
$z = t$	38	2

周辺確率と条件付き確率の関係

$$p(z = s)$$

- 上は、xの値が未知のとき、$z = s$である確率

$$p(z = s|x = b)$$

- 上は、$x = b$が観測されているとき、$z = s$である確率

- この２つの確率$p(z)$と$p(z|x)$の関係は？→ベイズ則

ベイズ則 Bayes rule

$$p(z|x) \propto p(x|z)p(z)$$

- 出来事の観測で確率が変わるという式
  - ある仮説が成り立つ確率$p(z)$は…
  - その仮説が成り立っていると尤もらしさが増す（尤もらしさが減る）出来事xが観測されると…
  - 高い(低い)値$p(z|x)$になる

ベイズ則の証明

- 条件付き確率の定義から$p(z|x) = \frac{p(x,z)}{p(x)}$
- 条件付き確率の定義から$p(x|z) = \frac{p(x,z)}{p(z)}$
  - よって$p(x,z) = p(x|z)p(z)$
- 組み合わせると$p(z|x) = \frac{p(x,z)}{p(x)} = \frac{p(x|z)p(z)}{p(x)} \propto p(x|z)p(z)$

ベイズ則の「比例する∝」という記号

$$p(z|x) \propto p(x|z)p(z)$$

- ∝は「左辺が右辺に比例する」という意味
- $p(z|x)$をちゃんと求めるには、比例定数を求める必要がある
- $\sum_z p(z|x) = 1$が満たされるように、比例定数を決める

$$p(z|x) = ???$$

ベイズ則を等号を使って書くと・・・

$$p(z|x) = \frac{p(x|z)p(z)}{\sum_z p(x|z)p(z)}$$

- 右辺の分母はxの周辺確率、つまり$p(x)$に等しい

$$\sum_z p(x|z)p(z) = \sum_z \frac{p(x,z)}{p(z)} p(z) = \sum_z p(x,z) = p(x)$$

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

例）ハンバーガー

- $p(z = s)$は、$z = s$であるとき尤もらしくなるデータが観測されると、高くなる

$$p(z = s|x = b) \propto p(x = b|z = s)p(z = s)$$

$$p(z = s|x = b) \propto \frac{8}{10} \times \frac{1}{5}$$

	$x = a$	$x = b$
$z = s$	2	8
$z = t$	38	2

- 「∝」は比例するという意味
- 比例関係をもとに$p(z = s|x = b)$の値を求めるには

  $\sum_z p(z|x = b) = 1$を使う（確率は、すべての場合について足すと１）

$p(z = s|x = b) \propto p(x = b|z = s)p(z = s)$より

$$p(z = s|x = b) \propto \frac{8}{10} \times \boxed{\frac{1}{5}} = \frac{4}{25}$$

$p(z = t|x = b) \propto p(x = b|z = t)p(z = t)$より

$$p(z = t|x = b) \propto \frac{2}{40} \times \frac{4}{5} = \frac{1}{25}$$

ということは$p(z = s|x = b)$の実際の値は

$$p(z = s|x = b) = \frac{\frac{4}{25}}{\frac{4}{25} + \frac{1}{25}} = \boxed{\frac{4}{5}}$$

	$x = a$	$x = b$
$z = s$	2	8
$z = t$	38	2

確率分布

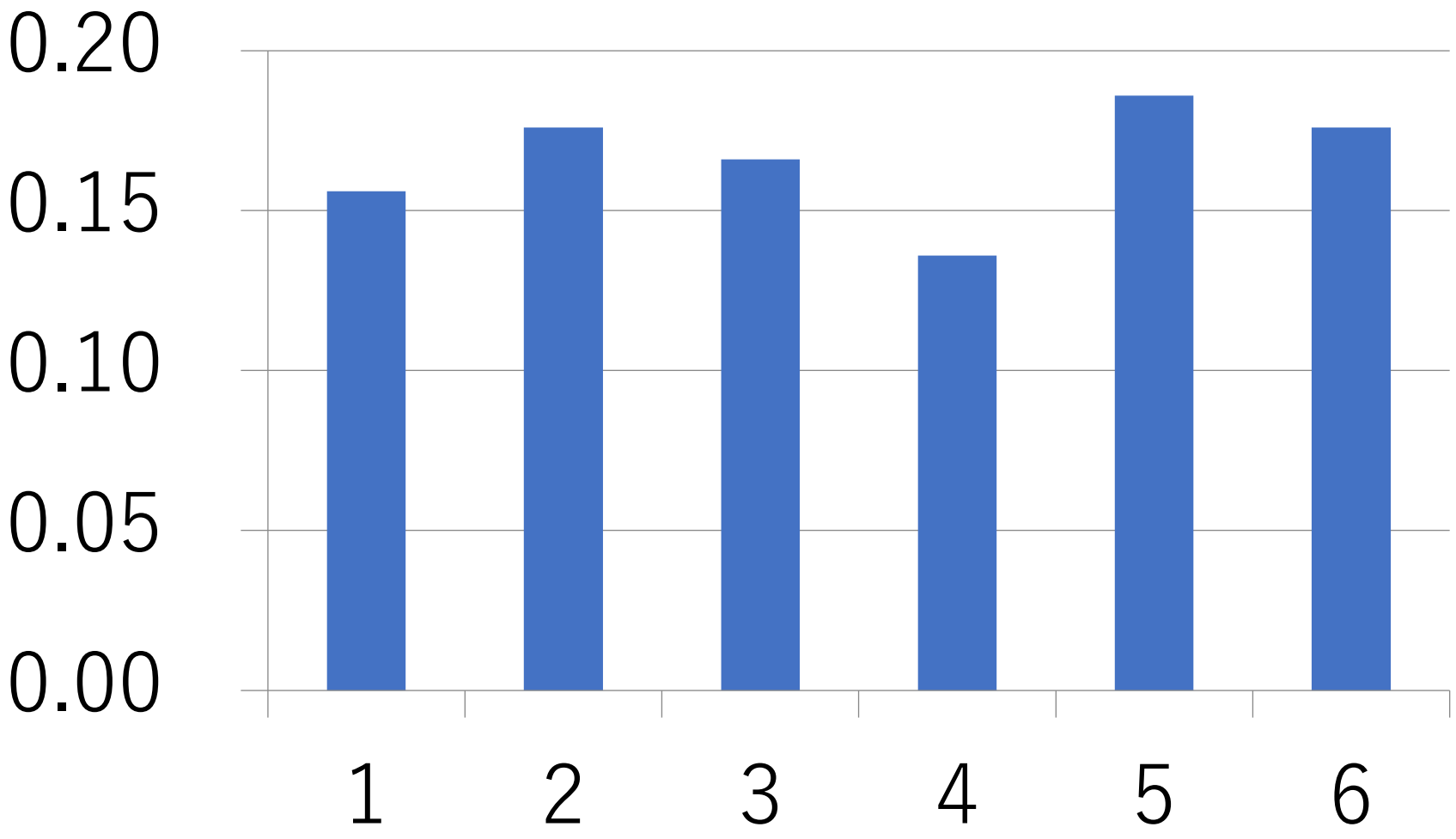
- 確率変数について、それがとりうる値のすべてについて、その値をとる確率を示したものを、確率分布と呼ぶ
- 確率変数がとりうる値すべてについて、その値をとる確率を足し合わせると、１になる
- 連続値をとる確率変数の場合は、和ではなく積分で考える

離散確率分布
discrete probability distribution

例）サイコロ

- 事象：「1の目が出る」，「2の目が出る」等
- 確率分布：各事象に確率を定めたもの
  - 離散的なので事象を一個一個と数えられる
  - すべての事象の確率を足すと1になる

例）サイコロの目の確率分布



連続確率分布
continuous probability distribution

例）体重測定

- 事象：「体重が57.5kg」，「体重が70.1kg」等
- 確率分布：事象の集合に確率を定めたもの
  - 連続的なので、事象を一個一個と個別に扱えない
  - すべての事象を含む集合の確率は1

正規分布

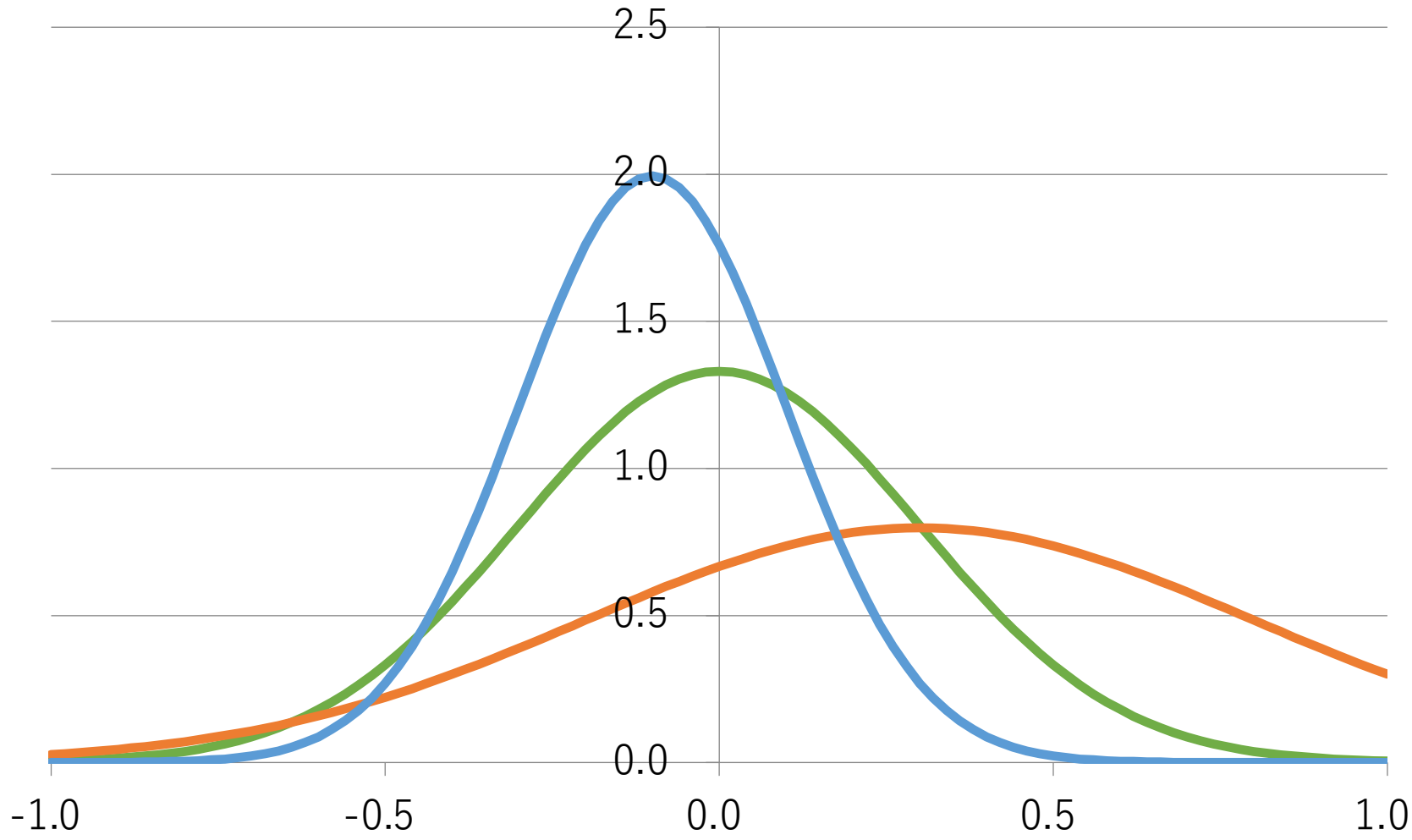
- 平均μと標準偏差σで決まる分布
- 正規分布をあらわす関数

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

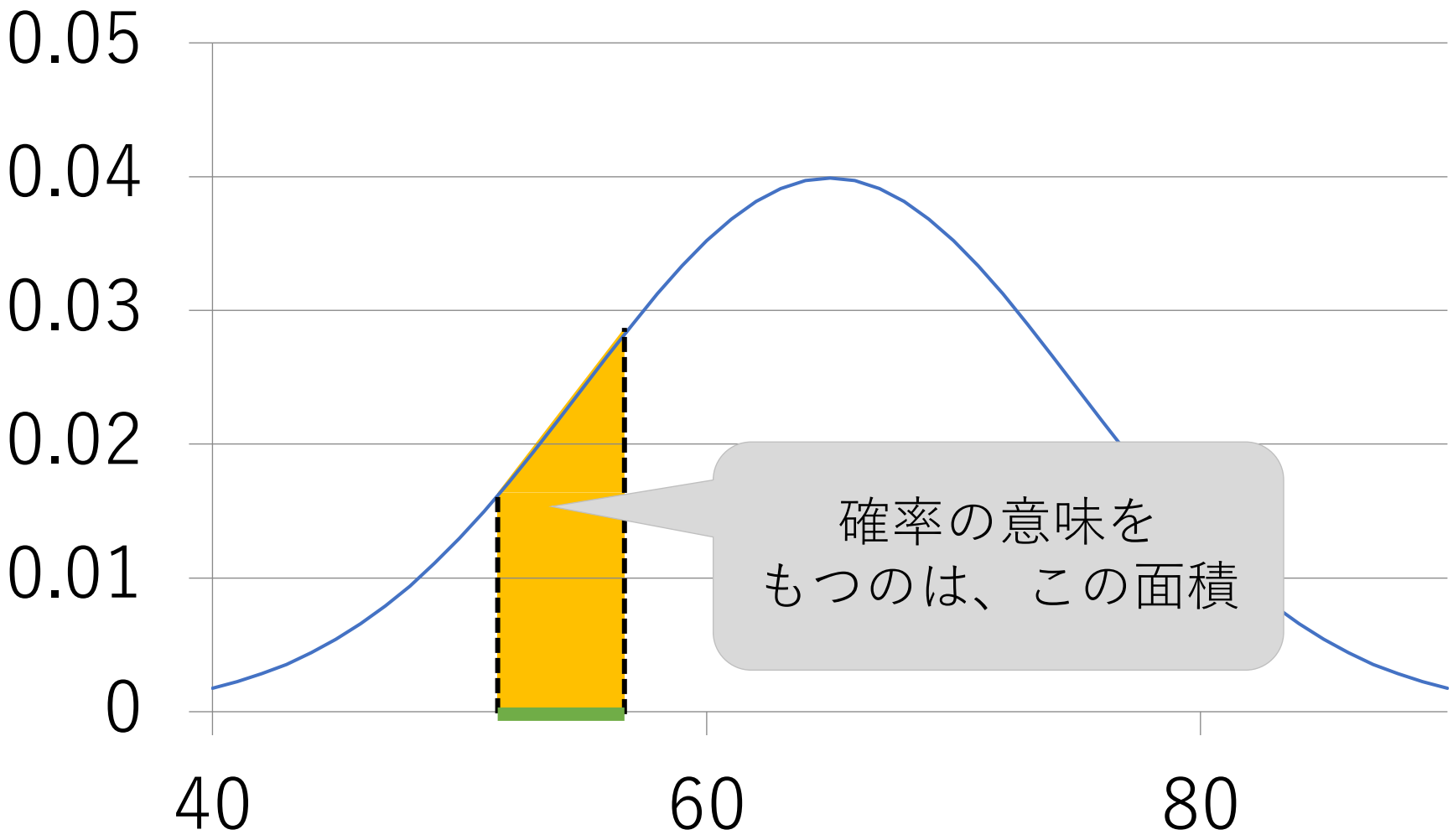
- これをある範囲で積分するとその範囲の値をとる確率が求まる

例）$\int_{45}^{55} f(x)dx$は「体重が45~55kg」という事象の確率

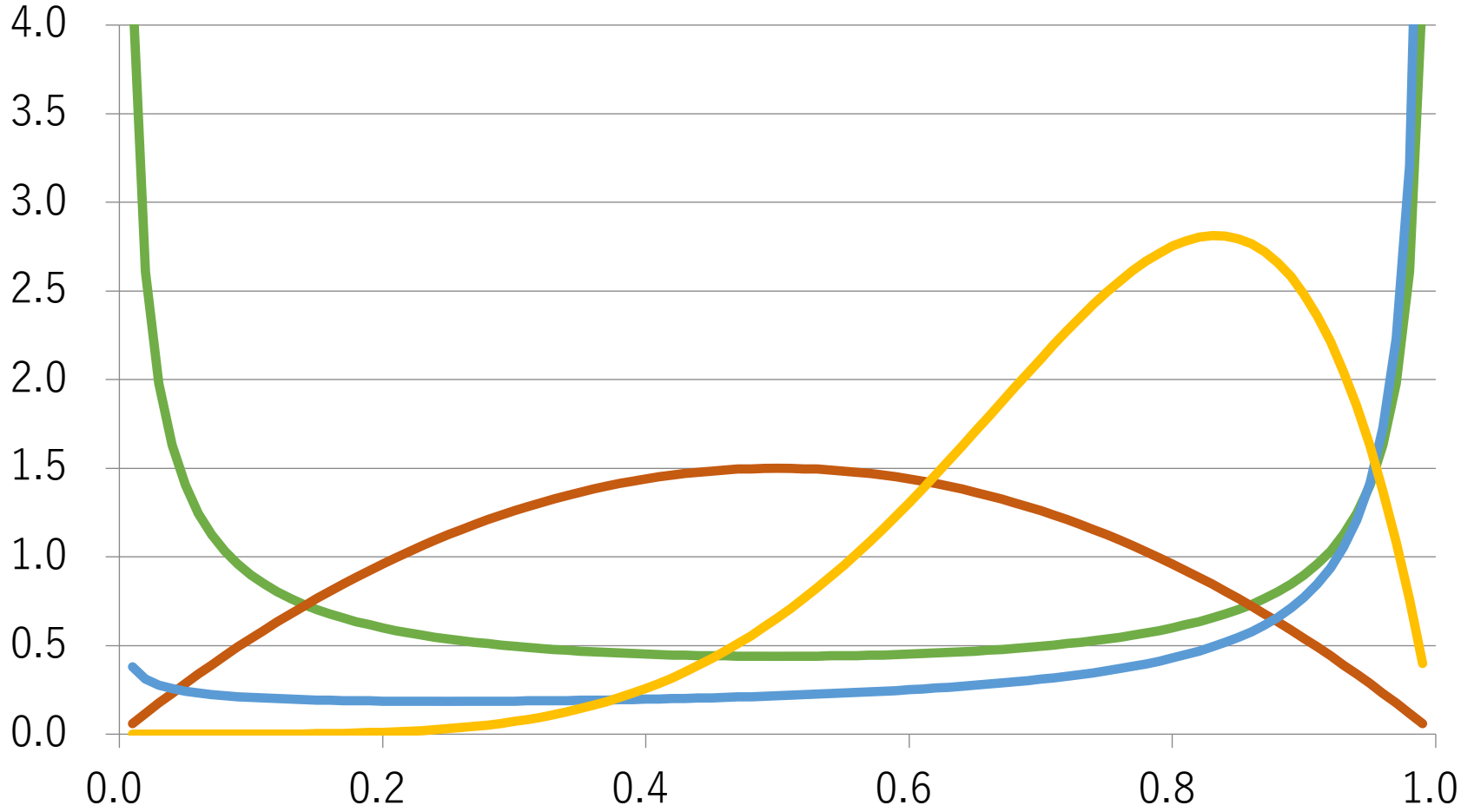
正規分布 normal distribution



例）体重の確率分布



ベータ分布 Beta distribution



期待値 (1/2)

- 確率変数xの期待値とは、その変数がとりうる値とその値をとる確率との積を、とりうる値すべてにわたって加算したもの
- 連続値をとる確率変数の場合は、加算するのではなく積分する

離散確率変数の場合：$\sum_x xp(x)$

連続確率変数の場合：$\int xp(x)dx$

期待値 (2/2)

- 変数xの関数$f(x)$についても、各々の値をとる確率を掛けて和や積分をとることで期待値が得られる

離散確率変数の場合：$\sum_x f(x)p(x)$

連続確率変数の場合：$\int f(x)p(x)dx$

  - 前のスライドは、$f(x)$が恒等関数の場合を説明していたとも言える

問題1-3

- 4枚のフェアなコインを投げたとき、表が出た枚数と裏が出た枚数の差の絶対値の期待値はいくら？
  - フェアなコインとは、表が出る確率も、裏が出る確率も、ぴったり0.5のコインのことをいう

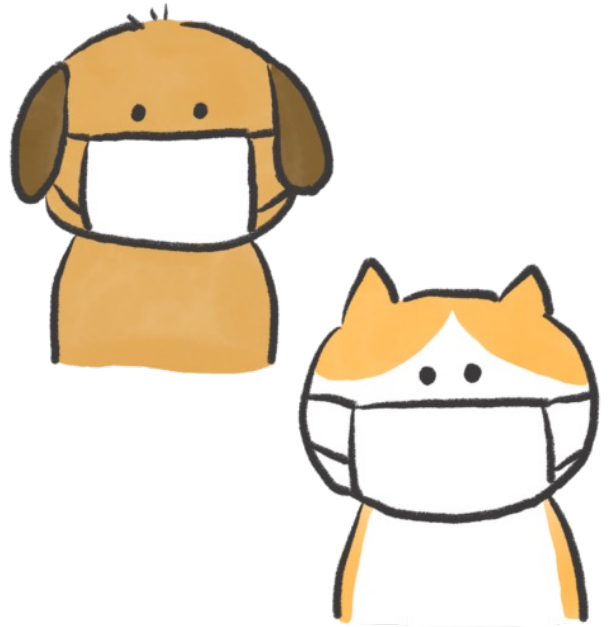
問題1-3の答え

$$\sum_{i=0}^{4} |i-(4-i)| \times \frac{4!}{i!\,(4-i)!} \times \left(\frac{1}{2}\right)^4$$

$$= \left(\frac{1}{2}\right)^4 \{4\times 1 + 2\times 4 + 0\times 6 + 2\times 4 + 4\times 1\} = \frac{24}{16} = \frac{3}{2}$$

本日の課題

- バスにわんこが20匹、にゃんこが15匹乗っている。
- わんこ20匹中、12匹が白い毛である。
- にゃんこ15匹中、4匹が白い毛である。



- このバスの中から無作為に1匹、乗客を選ぶ。

1. 選んだ乗客が白い毛である確率を求めよ。
2. 選んだ乗客がわんこであったときに、その乗客が白い毛である確率を求めよ。

https://fotoma.net/