

データサイエンスへの誘い

第7回: データと確率

瓜生真也 (デザイン型AI教育研究センター・助教)

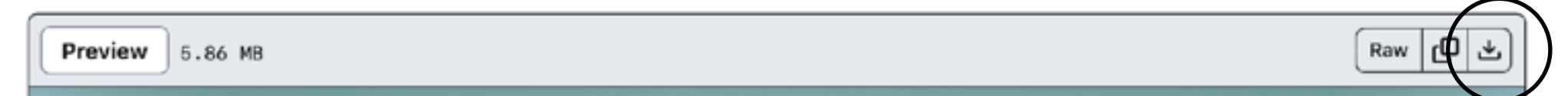
講義内容（予定）

講義に関する資料（スライド、補足資料等）を  GitHubに置いておきます

 <https://github.com/uribo/INNV1250>



ダウンロード可能



- | | |
|--|---|
| <ol style="list-style-type: none">1. ガイダンス、データサイエンスとは何か2. プログラミング基礎3. 再現可能性4. データ処理の手法5. データの要約6. データの可視化7. データと確率8. データからの推論 | <ol style="list-style-type: none">9. 複数のデータを比較する10. 統計のウソ11. 統計的モデリング12. 統計的学習13. さまざまなデータサイエンスの手法14. 機械学習と人工知能（AI）15. 期末試験16. 振り返りと統括 |
|--|---|

今日の目標

統計的推測と仮説検定の基礎となる

確率、確率分布を理解する

【課題】 確率・期待値についてのクイズ

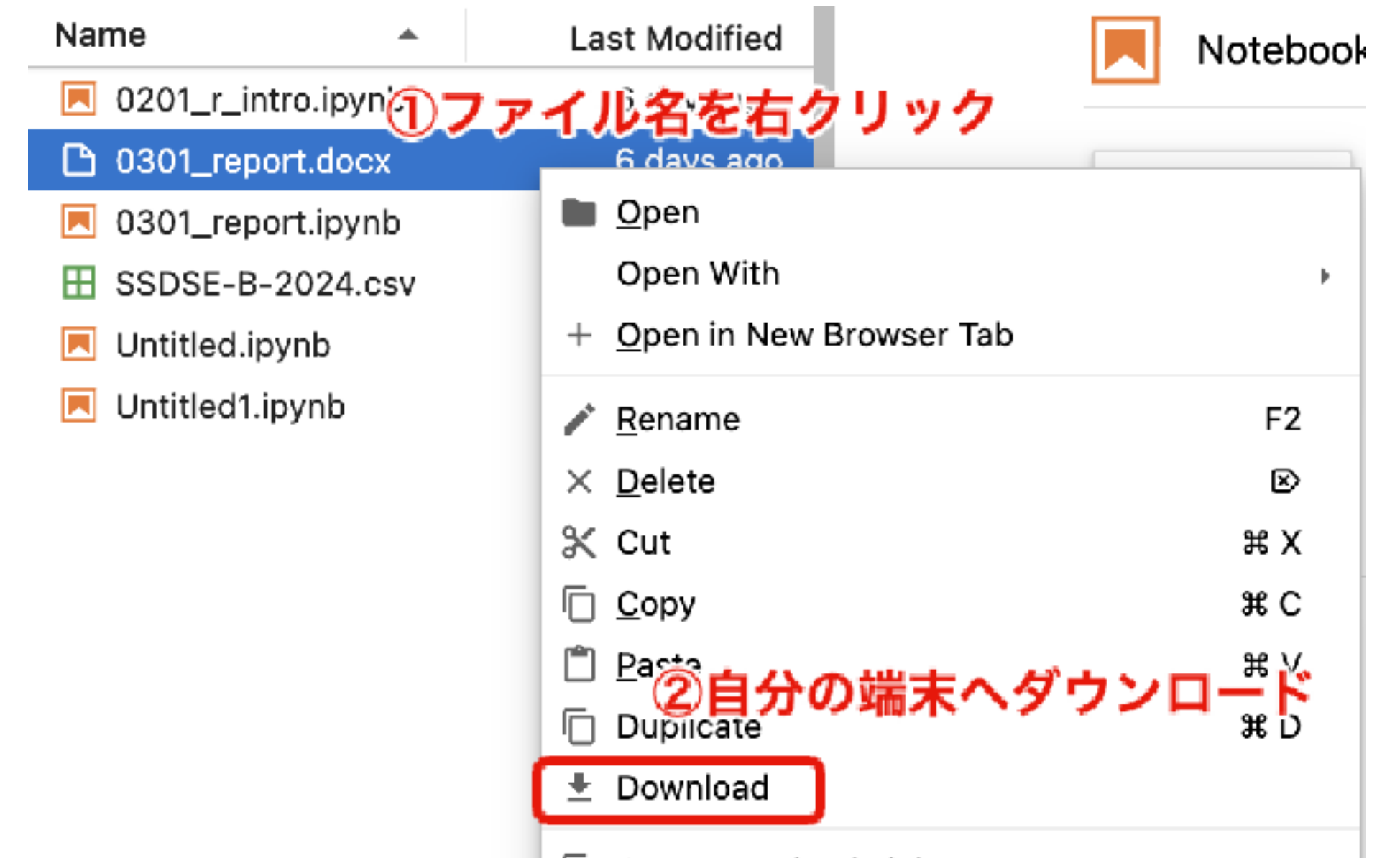
提出期限: 来週の講義開始前まで

手順

1. 添付ファイルをダウンロード
2. JupyterHubへアップロード
3. コードやコメントを記述、実行
4. 保存
5. ダウンロードしたファイルをmanabaへアップロード

manabaのレポートとして提出してください

メニュー上の「ファイル」から「ダウンロード」



注意: ファイル名は英数字のみにすること

日本語（漢字、片仮名、平仮名）、全角英数字、スペース、記号等は使わない

ファイルをダウンロードしても開けなくても問題ない（気にしない）

内容の確認、編集はJupyterHub上で行う

統計的推測

統計的推測

観測されたデータを用いて、未知の母集団の特性を得るための一連の手続き

分布は？平均値は？分散は？

統計的推定と**仮説検定**の2つの主要な手法を利用する

統計的推定

標本の特性から母集団の特性を推測
点推定と区間推定が含まれる

(よく使われる例え) 味噌汁の味を確認するための味見

仮説検定

データを用いて特定の仮説が
真であるかを評価する

確認された事象が偶然生じたものなのか

→いずれも「統計学」「確率」を利用する

まずは確率とその背景について理解を深めよう

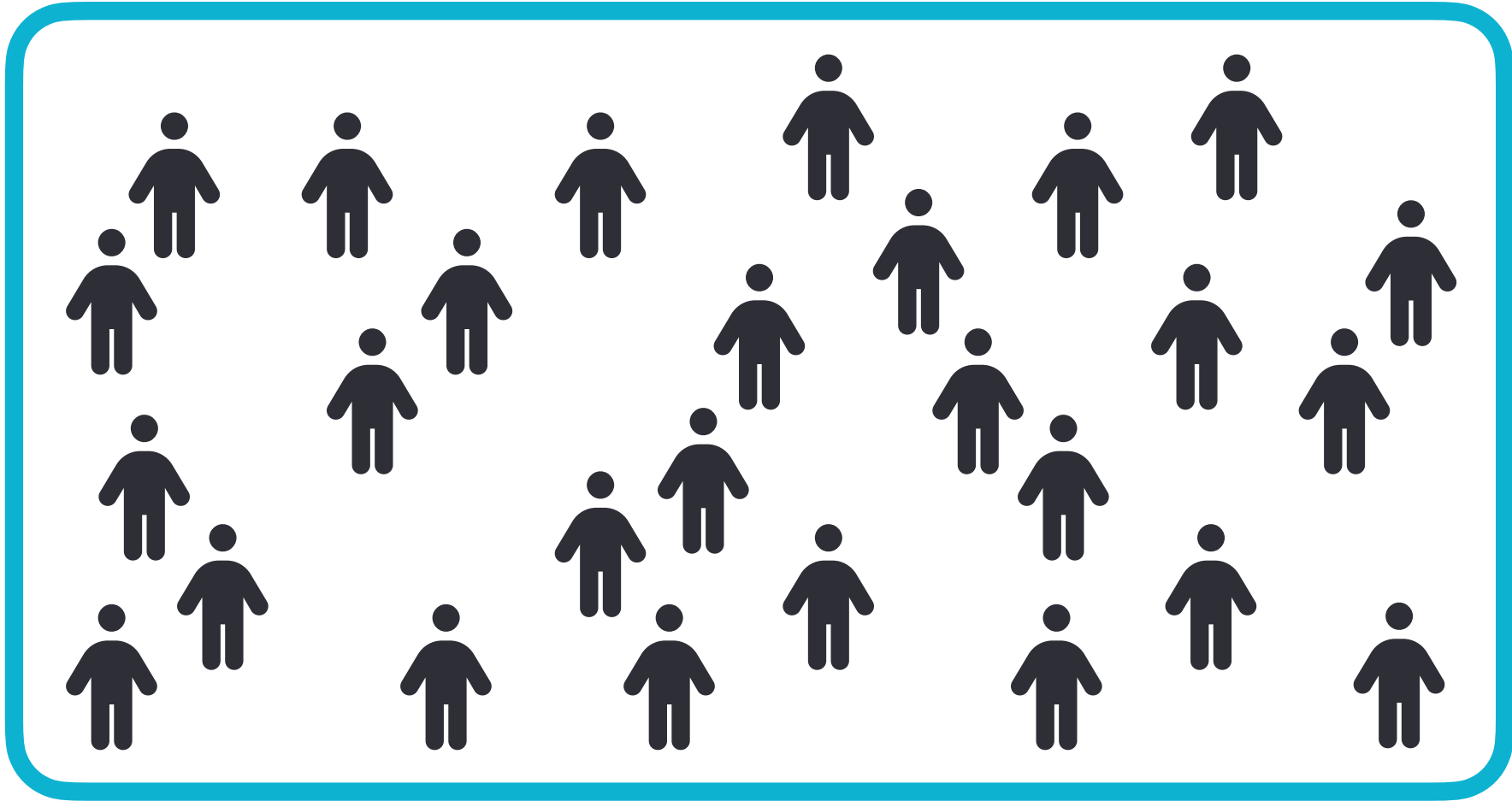
母集団と標本

関心のある特定の集団全体とその一部

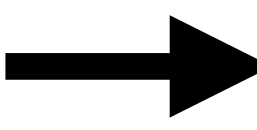
標本から母集団の特性を推定する

母集団

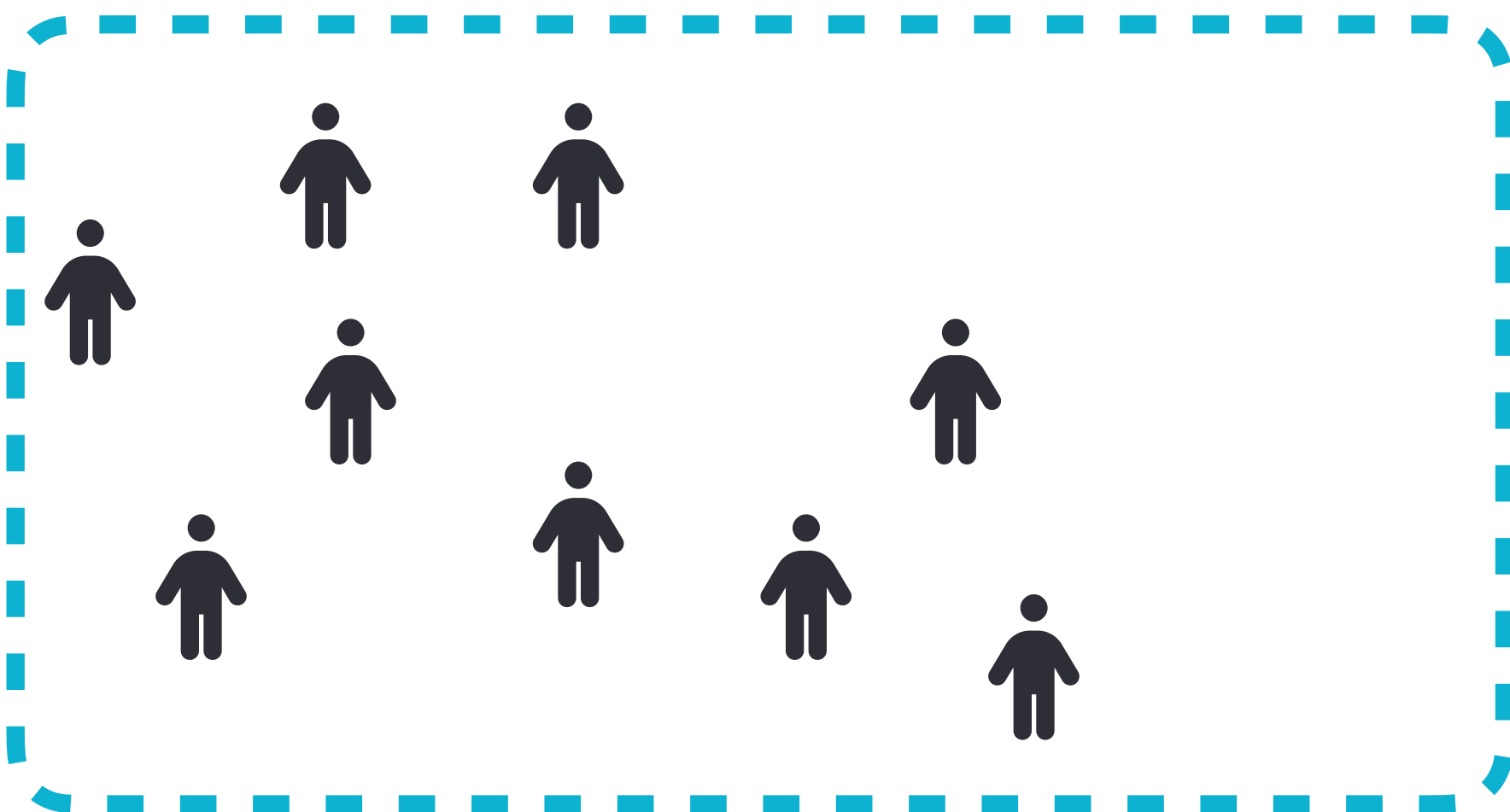
関心のある特定の集団全体



国民の全て、全ての製品、全ての患者など



標本調査

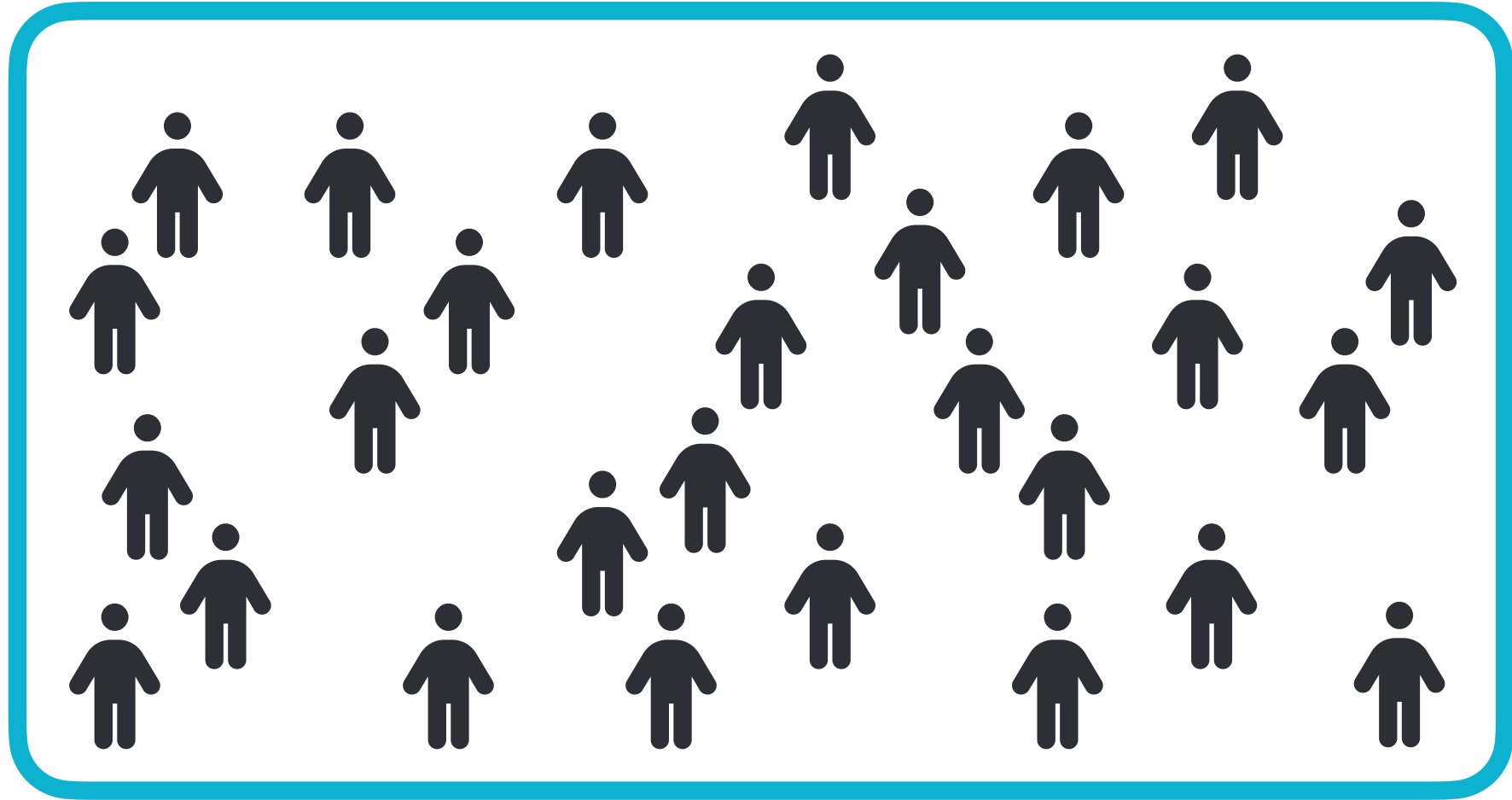


母集団から標本を抽出

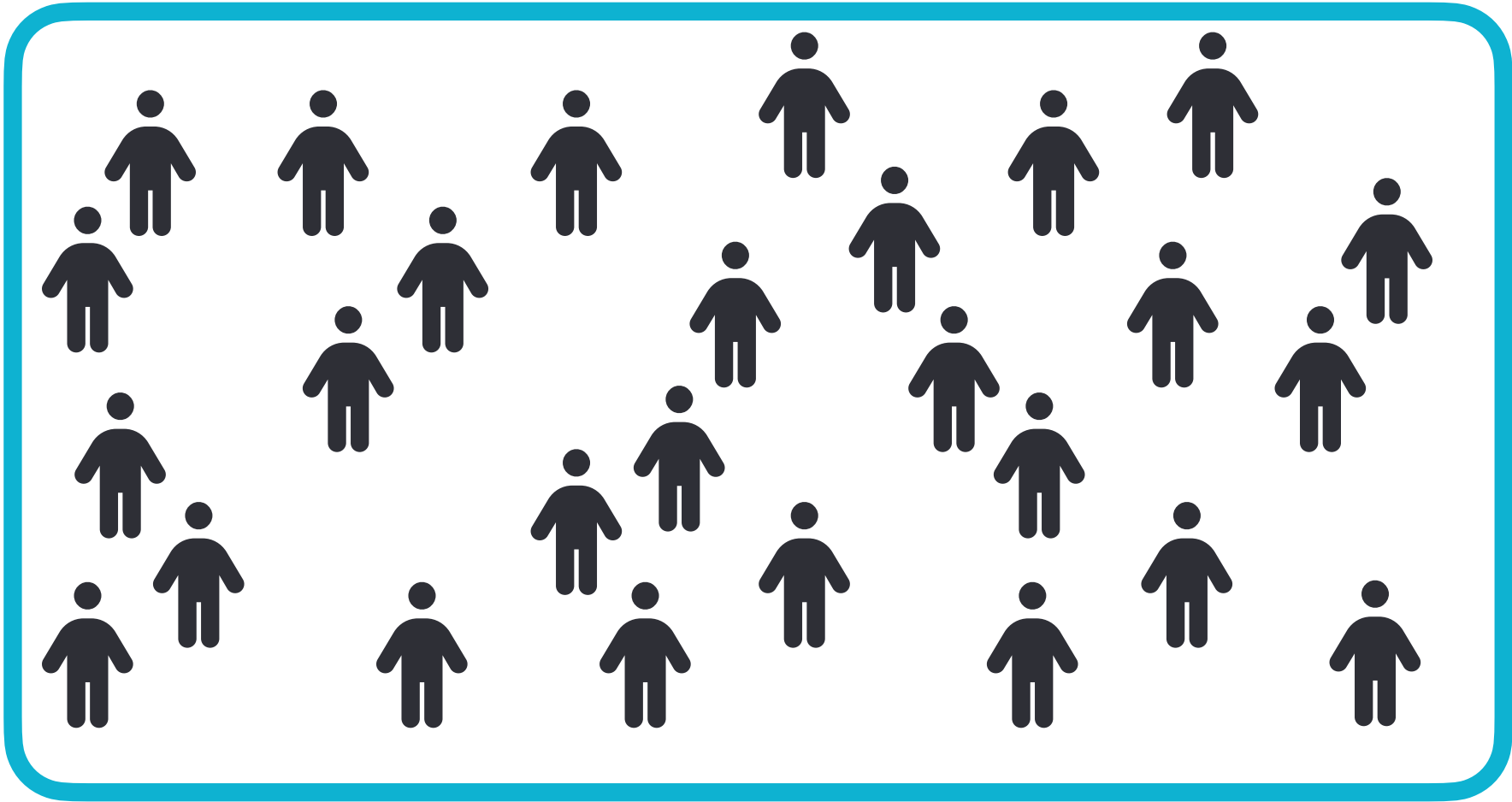
手間や費用の問題から行われる調査方法

母集団と標本

母集団

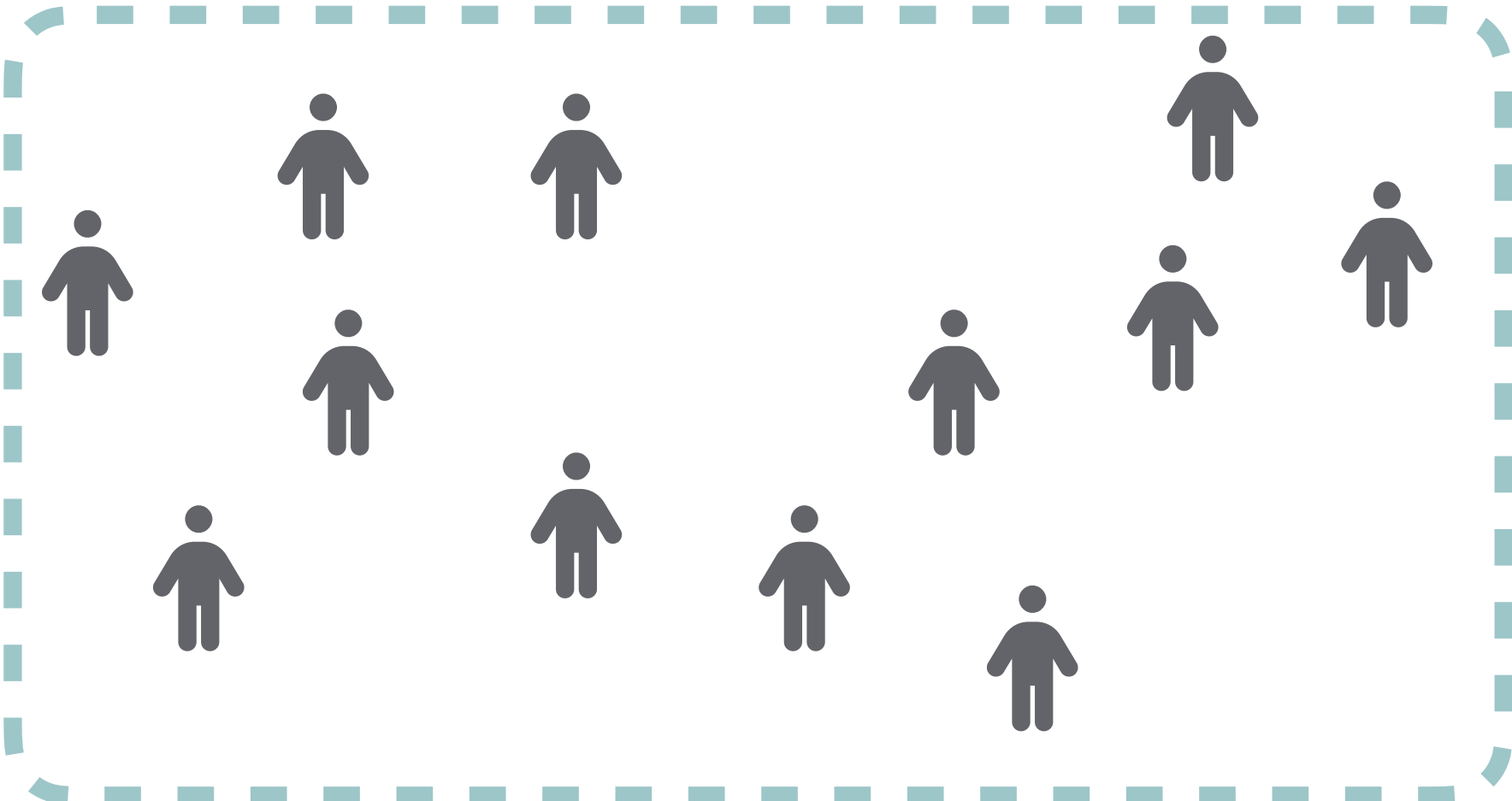


全数調査



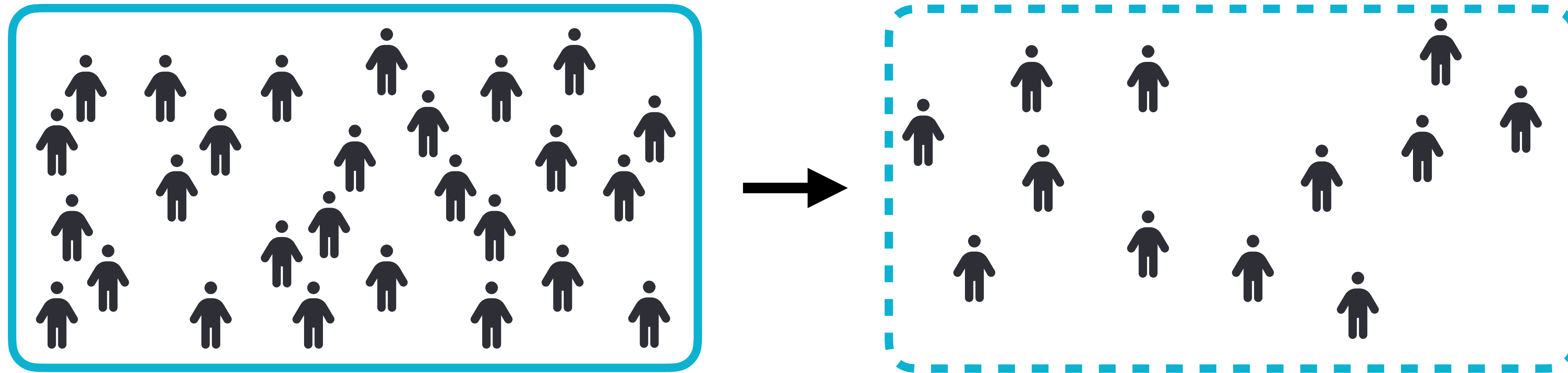
母集団全体を調査。国勢調査など
母集団の特性を反映したデータが得られる

標本調査



標本抽出: 無作為抽出法

母集団から無作為に標本を抽出

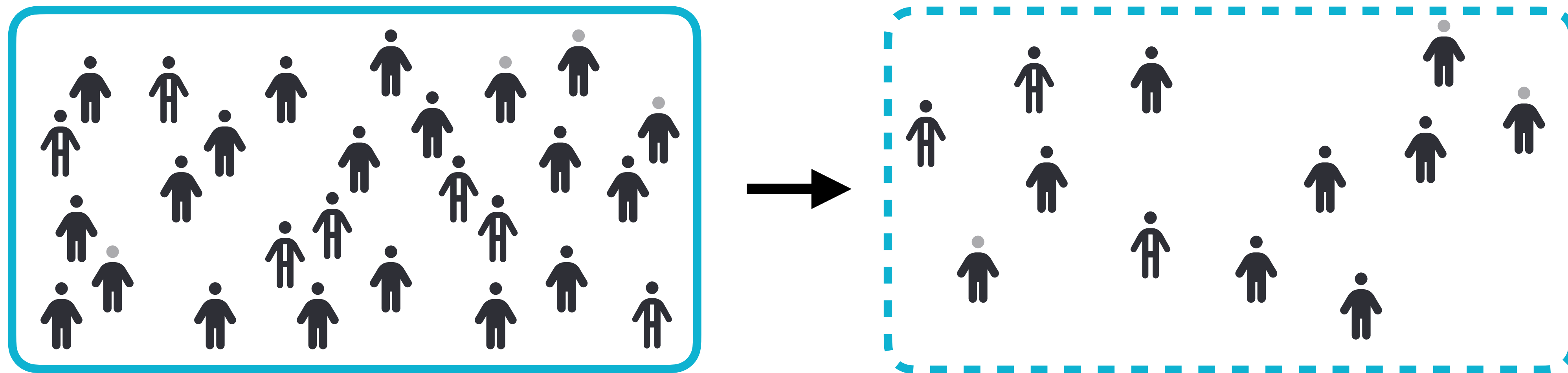


```
set.seed(123)
# 1から100までの整数を生成
population <- seq.int(100)
# 10個の標本を無作為に抽出
sample <- sample(population, 10)
sample
#> [1] 31 79 51 14 67 42 50 43 97 25
```



標本抽出: 層化抽出法

母集団を構成する要素の特徴に基づき、母集団からの抽出を行う



性別や年齢、職業などの特徴で母集団を分ける
→性別層、年齢層、職業層...
層ごとに無作為抽出
層を代表する標本を得る

```
library(dplyr)
# irisのspeciesから10件ずつ抽出
iris %>%
  slice_sample(n = 10,
               replace = FALSE,
               by = Species)
```



確率と確率変数

確率

観測される結果が偶然によって決まる事象を表す

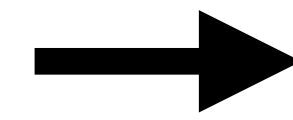


コイン投げ

候補

{表, 裏}

一度行う



{表}

確率

$\frac{1}{2}$



サイコロ投げ

{1, 2, 3, 4, 5, 6}



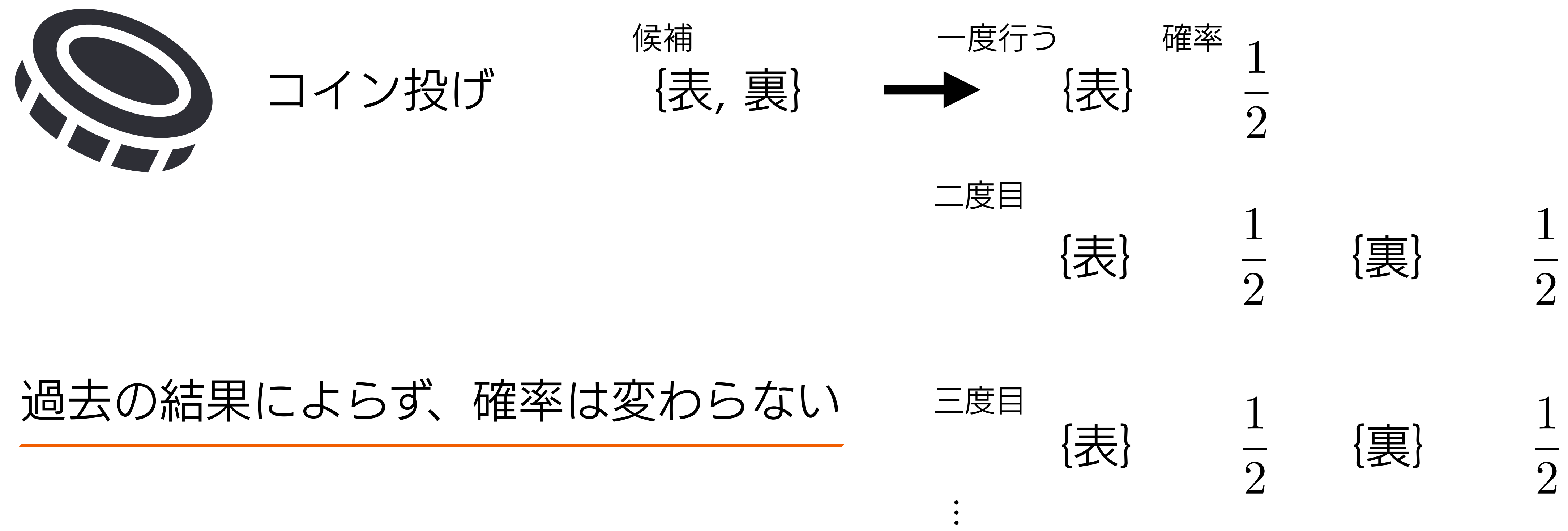
{1}

$\frac{1}{6}$

→ 事象の「起きやすさ」 を数値で表現する

事象の独立性

複数の事象が互いに影響を与えずに発生する性質



確率変数

ある変数のとる値を確率Pに従って出現するとみなす→事象を数値的に扱えるように

 コイン投げ

表なら1

裏なら0

$$P(X = 1) = \frac{1}{2} \quad P(X = 0) = \frac{1}{2}$$

すべての可能な出力値についての確率の和は1となる

離散型確率変数

→有限個の値をとる変数

コインの裏表、サイコロの出目など

連続型確率変数


→ある範囲の値をとる変数

身長や降水量など

離散・連続の違いは第4回の内容参照

確率分布

確率変数を取り得る値とその値が出現する確率との対応関係を表す

 コイン投げ

表なら1 裏なら0

離散型確率変数

事象	X	$P(X)$
表	1	1/2
裏	0	1/2

一様分布… すべての事象（結果）が同じ確率で発生する確率分布

確率変数の特定の値が取る確率を関数で表現

確率関数

離散型確率変数 X が取り得る値すべてに確率を割り当てる

$$P(X = x_i) = p_i, \quad i = 1, 2, \dots, m$$

- すべての可能な出力値 x について、確率は0以上
- すべての可能な出力値についての確率の和は1

$$\sum_{i=1}^m P(X = x_i) = 1$$

→ 確率の公理に従う

確率分布の形: 二項分布

🎯 ベルヌーイ試行（結果が2通りにしかない確率実験）をn回行ったときに
特定の事象（例えば成功）が出現する確率分布

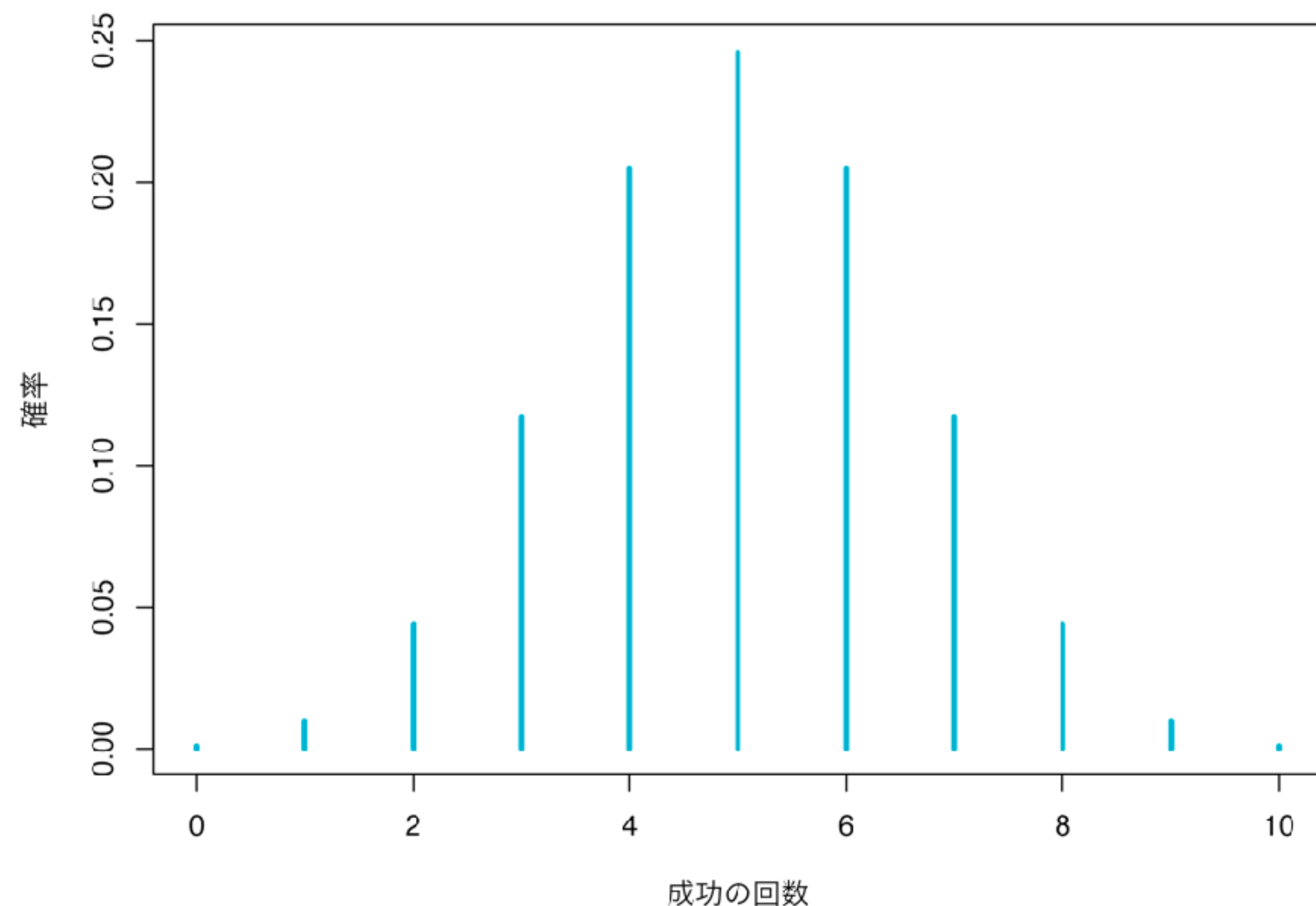
確率関数 $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ $\binom{n}{k} = \frac{n!}{k!(n-k)!} = {}_n C_k$

n 試行回数

p 1回の試行での成功確率

k 試行回数nのうちでの成功の回数

二項分布（試行回数=10成功確率=0.5）



```
# 成功確率
p <- 0.5
# 試行回数
n <- 10
# 二項分布の確率関数を計算する範囲
x <- 0:n
# 二項分布の確率関数を計算
prob <- dbinom(x, size = n, prob = p)
# グラフ描画
plot(x, prob, type = "h",
      ylab = "確率",
      xlab = "成功の回数",
      lwd = 3,
      col = "#0cb3d1")
```

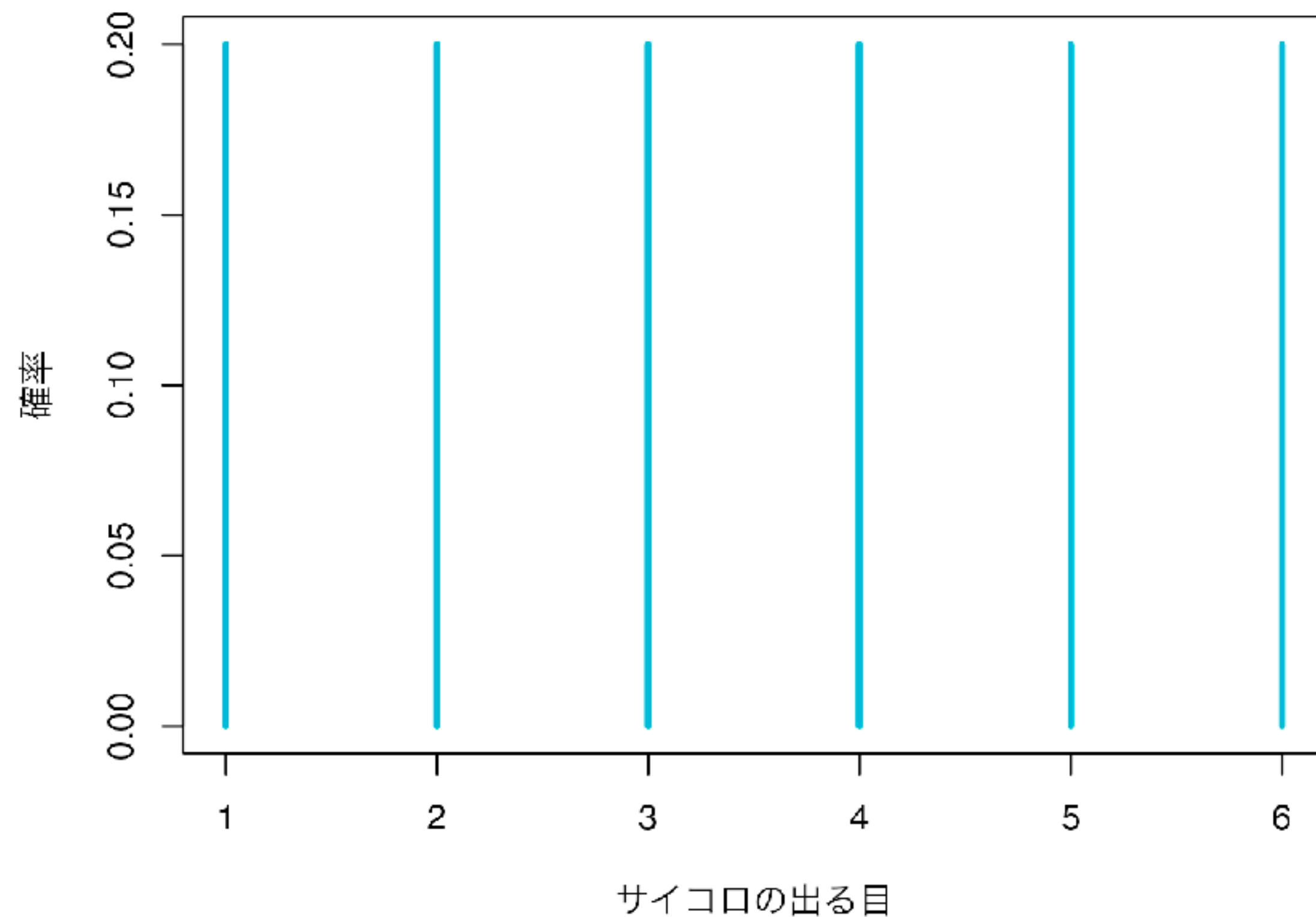


確率分布の形: 一様分布

サイコロの出る目 $\{1, 2, 3, 4, 5, 6\}$ … 離散一様分布

確率関数 $P(X = x_i) = \frac{1}{m}$, $i = 1, 2, \dots, m$ すべての事象が同じ確率で発生する

サイコロの出る目の離散一様分布



```
x <- seq.int(6)
y <- dunif(x, 1, 6)
plot(x, y,
     type = "h",
     lwd = 3,
     col = "#0cb3d1",
     ylim = c(0, 0.2),
     xlab = "サイコロの出る目",
     ylab = "確率",
     main = "サイコロの出る目の離散一様分布")
```



ここまでの内容の整理

- ・ 関心のある事象の「起きやすさ」を数学的に扱うために確率が利用される
- ・ ある確率をともなっていて生じる変数のことを確率変数 X と呼ぶ

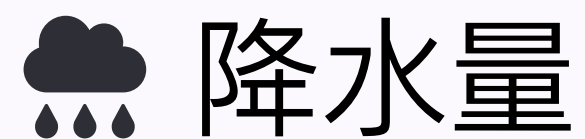
離散型確率変数

→有限個の値をとる変数

連続型確率変数

→ある範囲の値をとる変数

- ・ 確率変数が取り得る値とその確率との対応関係を確率分布によって表す
 - ・ 離散型確率変数では確率関数によって確率を割り当てる
 - ・ 確率分布の理解には、グラフへの描画が有効



連続型確率変数

連続型変数の確率分布

(理論上は無限に取り得る値が存在する) をどう表現する？

確率変数の特定の値が取る確率を関数で表現

確率密度関数 特定の範囲内に連続型確率変数の値が存在する確率を表現する

連続型確率変数は理論上無限の値を取る可能性があるため、特定の値が取る確率を求めると0

$$f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x) dx = 1$$

- 確率密度関数は非負の値。

$$f(x) \geq 0$$

どのような x の値に対しても $f(x)$ は0または0より大きい

- 確率密度関数により得られる確率の和は1に等しい
- 確率密度変数が定義する全範囲を積分した結果は1

$$\int f(x) dx = 1$$

→ 確率の公理に従う

確率密度関数の曲線下の面積（積分値）を通じて確率を扱う

確率分布の形: 正規分布

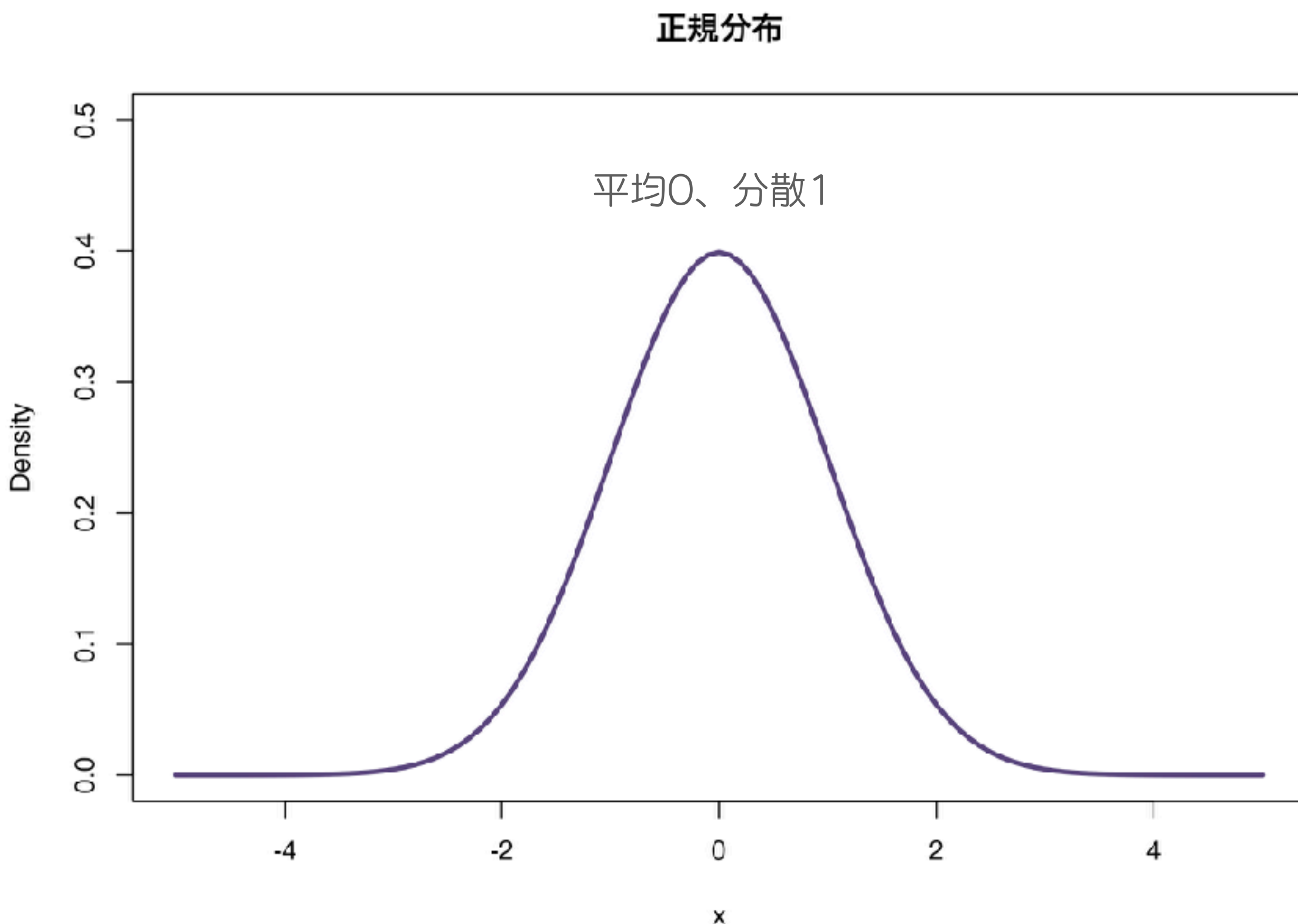
分散に対して平方根を求めたものが標準偏差

身長や体重など (連続型確率変数)

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

μ 平均値

σ^2 分散



xの範囲を設定

```
x <- seq(-5, 5, by = 0.1)
```

```
y1 <- dnorm(x, mean = 0, sd = 1)
```

グラフを描画

```
plot(x, y1, type = "l",
```

```
lwd = 3,
```

```
col = "#57467b",
```

```
ylim = c(0, 0.5),
```

```
xlab = "x",
```

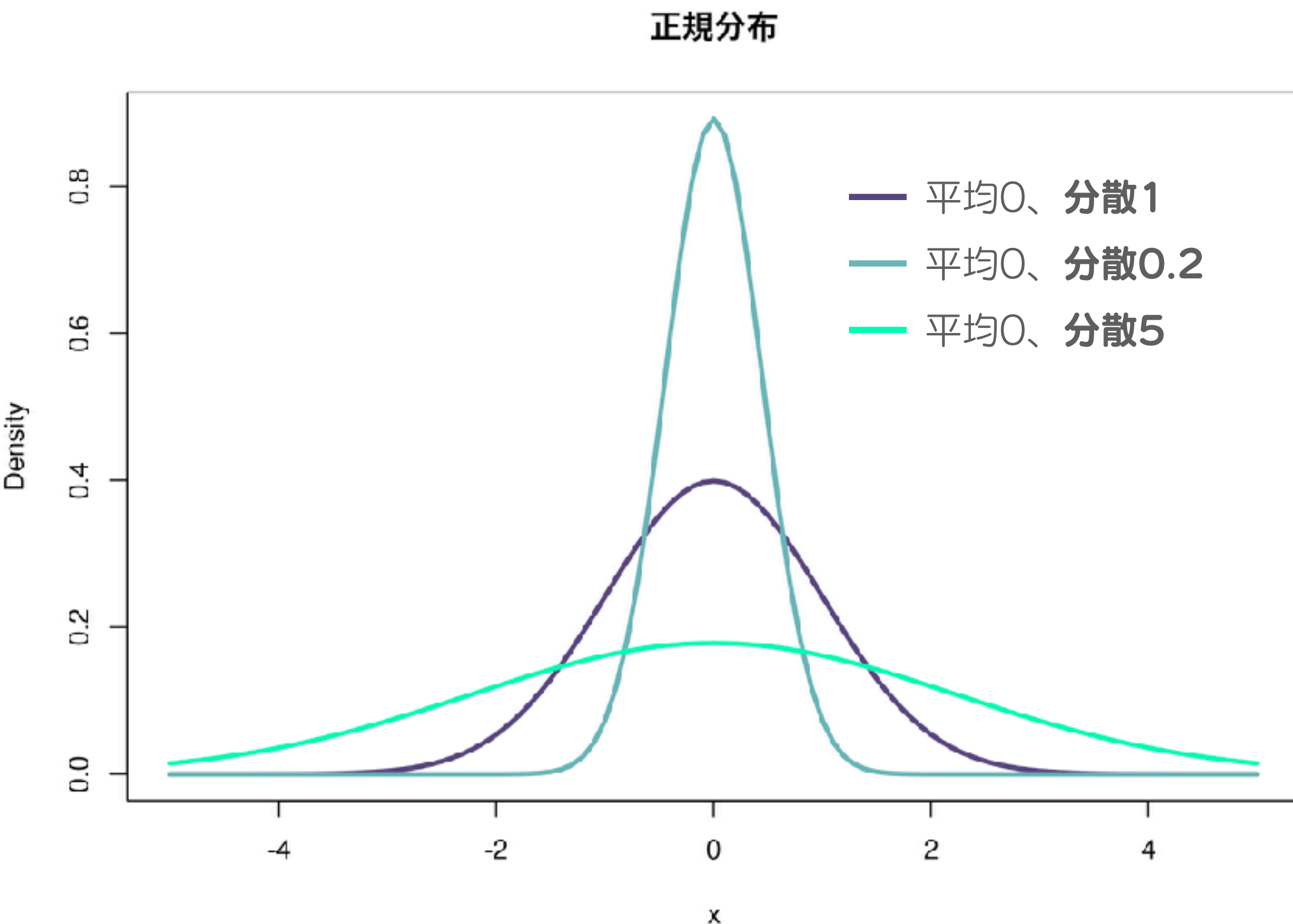
```
ylab = "Density",
```

```
main = "正規分布")
```



確率分布の形: 正規分布

正規分布の形状は平均と分散の2つのパラメータで決まる



```
# 平均0で分散がそれぞれ0.2、5の
# 正規分布の確率密度関数を計算
y2 <- dnorm(x, mean = 0, sd = sqrt(0.2))
y3 <- dnorm(x, mean = 0, sd = sqrt(5))
plot(x, y1, type = "l",
      lwd = 3,
      col = "#57467b",
      ylim = c(0, max(c(y1, y2, y3))),
      xlab = "x",
      ylab = "Density",
      main = "正規分布")
lines(x, y2, lwd = 3, col = "#7cb4b8")
lines(x, y3, lwd = 3, col = "#70f8ba")
```



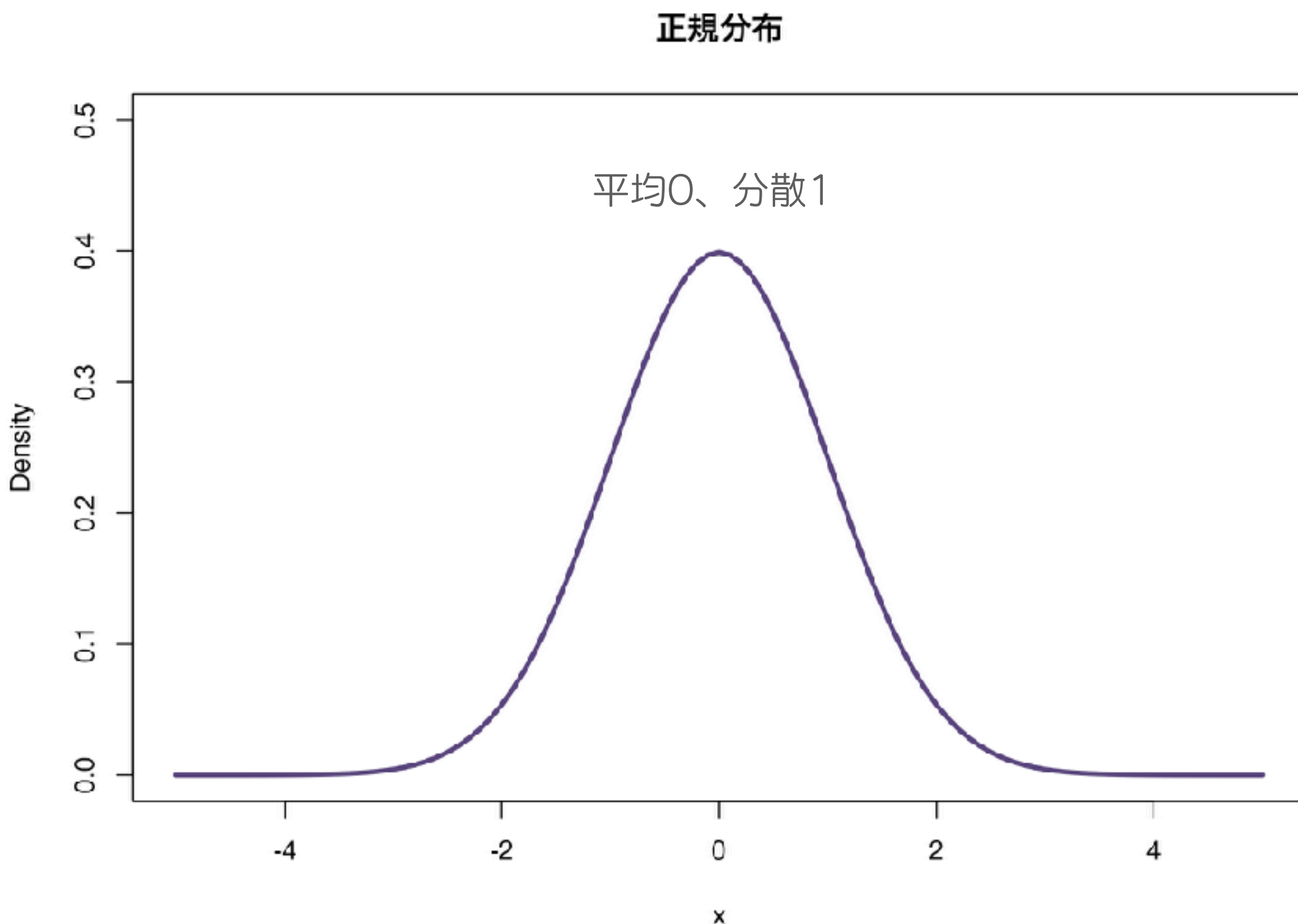
標準正規分布

平均値が0、標準偏差が1の正規分布

標準化により、ことなる正規分布を比較しやすくなる

確率を計算するための「標準正規分布表」が用意されている。

→特定の値や範囲に対する確率（面積）の計算が容易



標準化

期待値（平均）0、分散1の
確率変数に変換

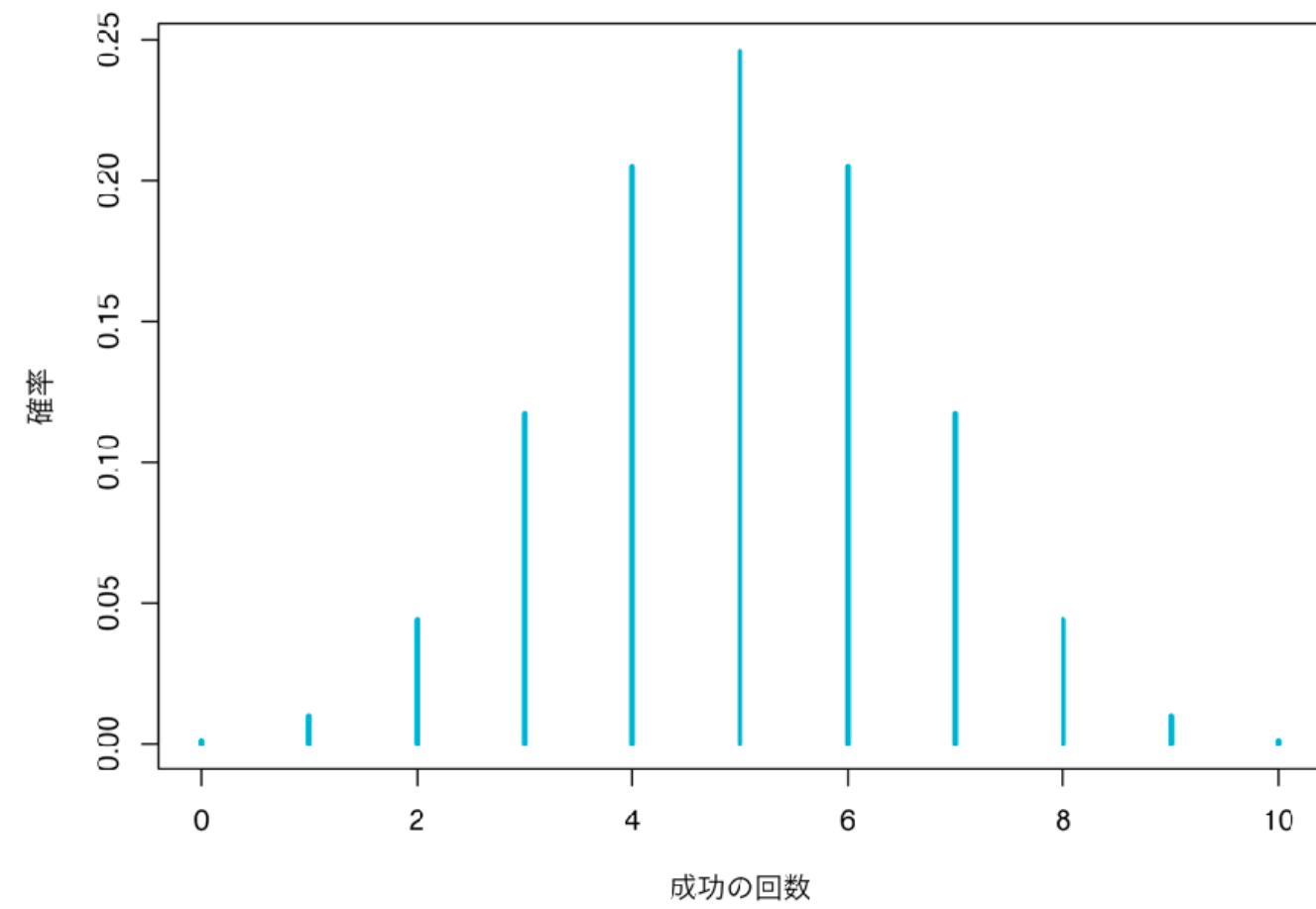
$$Z = \frac{x - \mu}{\sigma}$$

 `scale(iris$Sepal.Length)`

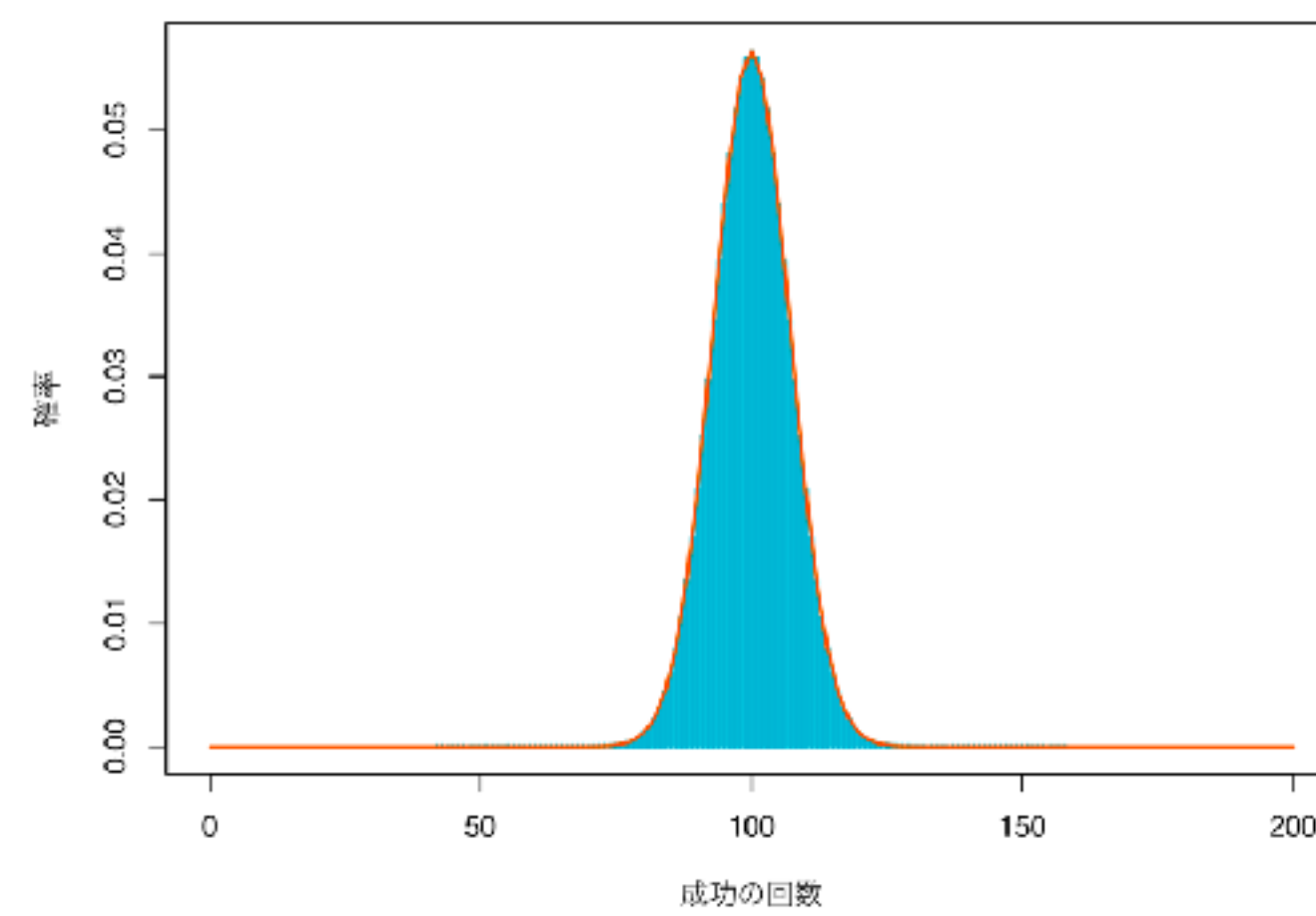
中心極限定理

独立かつ同一分布に従う確率変数の和は、
サンプル数が無限大であれば標準正規分布に従う

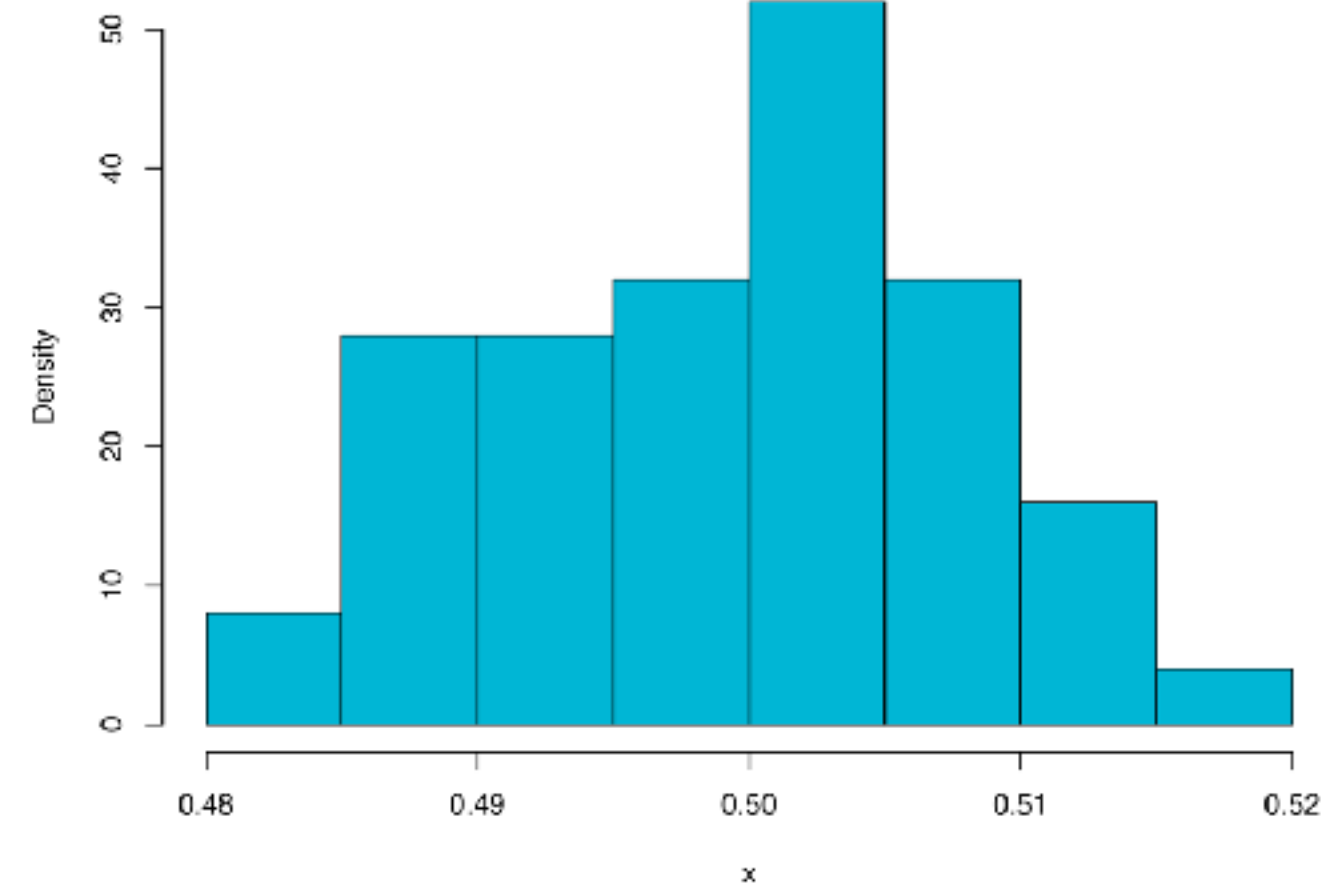
二項分布（試行回数 =10成功確率 =0.5）



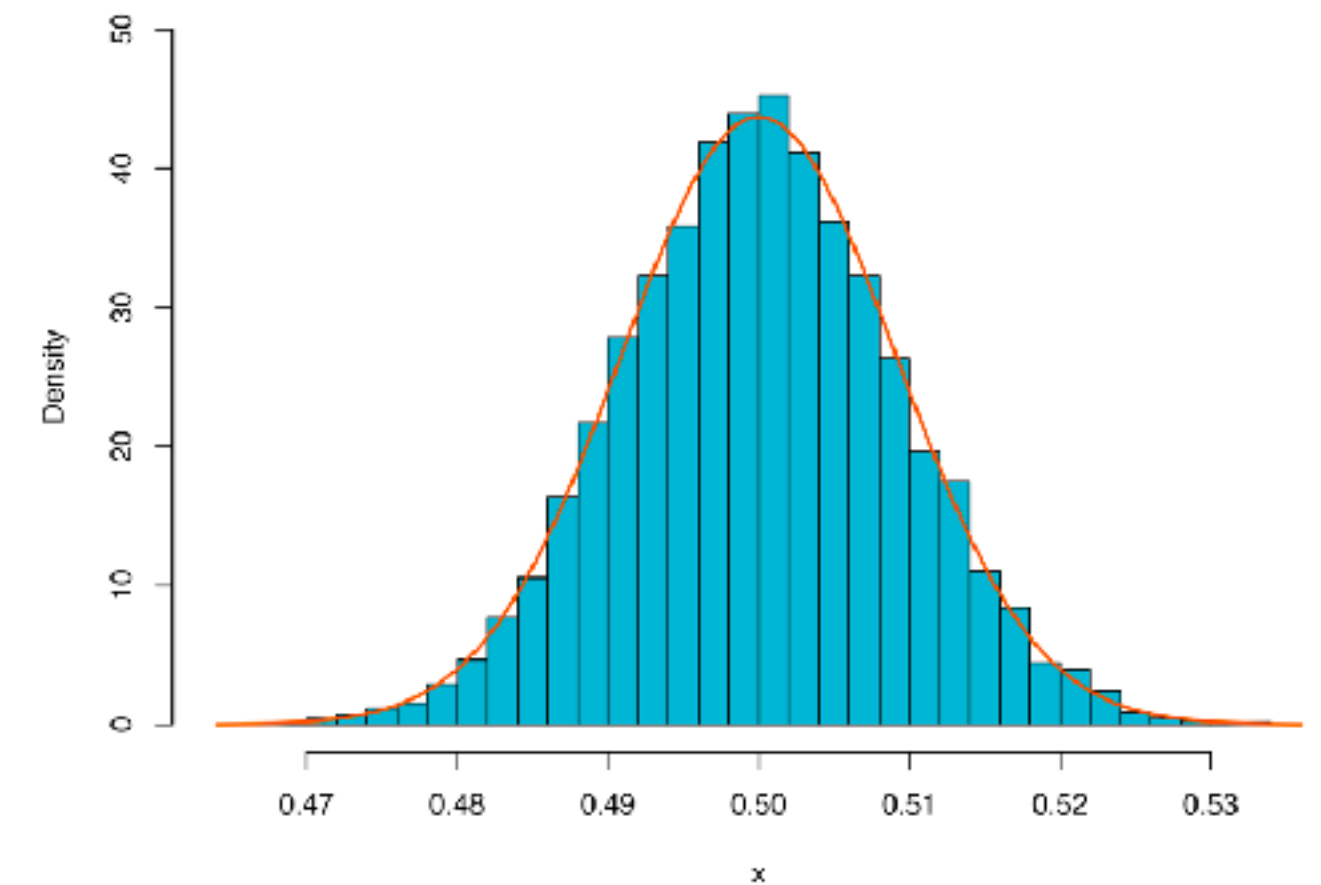
二項分布のグラフに正規分布の曲線を重ねる



一様分布



一様分布のグラフに正規分布の曲線を重ねる





期待値と分散

期待値: 確率分布の"中心"または"平均"の位置を示す

各事象がその出現確率によって「重み付け」された値（加重平均）の和

離散型確率変数

$$E(X) = \sum_x x P(X = x)$$

サイコロ投げの例（離散一様分布）  × 1/6  × 1/6  × 1/6  × 1/6  × 1/6  × 1/6

すべて足すと 3.5 → サイコロ投げを何度も繰り返し行くと平均的に3.5となる

実際に確率変数Xで取り得る値ではないことがある

連続型確率変数

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

特定の値が取る確率を求めると0

確率密度関数によって表される範囲を
確率として扱う

分散: 各値が期待値からどれだけ離れるかの平均を示す

第四回で扱った分散とは計算方法が異なるので注意

各結果の期待値からの差の二乗とその確率の積の合計

$$V[X] = E[(X - E[X])^2]$$

離散型確率変数

$$= \sum_i (x_i - \mu)^2 p_i$$

連続型確率変数

$$= \int (x - \mu)^2 \times f(x) dx$$

まとめ

第6回: データと確率のまとめ

- 「統計」および「確率」はデータサイエンスにおいて核となる概念である。
関心のある母集団について、標本から母集団の特性を知るための手続き
(統計的推測) が行われる→推定と仮説検定
- データは確率によって数値的に表現される
- 確率変数を取り得る値とその値が出現する確率との対応関係を表す確率分布には
さまざまな種類があり、パラメーターによってその分布の形が異なる。
- データサイエンスの手法では、データが正規分布に従うことを仮定したものが多い
正規分布の特性を利用する

参考資料・URL

目 小林正弘, 田畑耕治 『確率と統計』 (2021) 共立出版. ISBN: 978-4-320-11392-3
瓜生居室: なし、徳大図書館: あり、市立図書館: なし、県立図書館: あり

目 松井秀俊, 小泉和之(著), 竹村彰通 (編) 『統計モデルと推測』 (2019) 講談社.
ISBN: 978-4-06-517802-7
瓜生居室: あり、徳大図書館: なし、市立図書館: なし、県立図書館: なし

目 東京大学教養学部統計学教室 (編) 『基礎統計学I: 統計学入門』 (1991)
東京大学出版会. ISBN: 4-13-042065-8
瓜生居室: あり、徳大図書館: あり、市立図書館: なし、県立図書館: あり

目 東京大学教養学部統計学教室 (編) 『基礎統計学II: 人文・社会学の統計学』 (1994)
東京大学出版会. ISBN: 4-13-042066-6
瓜生居室: あり、徳大図書館: あり、市立図書館: なし、県立図書館: あり

目 滋賀大学データサイエンス学部, 長崎大学情報データ科学部 (編) 『データサイエンスの歩き方』
(2022) 学術図書出版社. ISBN: 978-4-7806-0936-3
瓜生居室: あり (電子版)、徳大図書館: あり、市立図書館: なし、県立図書館: なし

